Neural Computing and Applications (2023) 35:4145–4182
https://doi.org/10.1007/s00521-022-07937-4

ORIGINAL ARTICLE

Toward human activity recognition: a survey

Gulshan Saleem[1] · Usama Ijaz Bajwa[1] · Rana Hammad Raza[2]

Received: 6 March 2021 / Accepted: 10 October 2022 / Published online: 20 October 2022

Abstract

Human activity recognition (HAR) is a complex and multifaceted problem. The research community has reported numerous approaches to perform HAR. Along with HAR approaches, various surveys have revealed HAR trends in various environments and applications. HAR is linked to a variety of technology-dependent daily life systems, such as human–computer interaction systems, security surveillance, video surveillance, healthcare surveillance, robotics, content-based information retrieval, and monitoring systems. Because of technological advancements, HAR trends change quickly and necessitate an up-to-date and broader perspective. This study offers an HAR taxonomy, which includes online/offline HAR, multimodal/unimodal HAR, handcrafted feature-based, and learning-based approaches. This study attempts to present the multidisciplinary nature of HAR, such as application areas, activity types, task complexities, benchmark datasets, and/methods. This research includes a comparative analysis of state-of-the-art HAR methods and a discussion of popular datasets. The selected studies have been categorized using taxonomy, and different attributes such as activity complexity, dataset size, and recognition rate have been used for their analysis. The comparative analysis of HAR approaches has also helped to highlight domain challenges and open research directions for HAR researchers to follow.

Keywords Activity recognition · Action recognition · Video datasets · Deep learning · Handcrafted features · Video analysis · Computer vision

1 Introduction

Human activity recognition (HAR) is used to detect and classify human activities under appropriate labels. Human activities are complex and evolve temporally, necessitating suitable division into sub activities, as illustrated in Fig. 1. Human activity is an ongoing task composed of single or multiple gestures, actions, and interactions. Gesture refers to the movement of body parts to emphasize speech, whereas action refers to the collective movement of body parts to complete a task. For example, moving head in negation is a gesture, walking is an action, and speaking loudly with unpleasant facial expressions is an angry behavior. Interaction is a collection of actions usually performed by two or more subjects, for example, a two-person conversation, fighting, cooking food, data entry, and car washing, etc. Group activities are performed by multiple persons and may include a collection of gestures, actions, and interactions, for example, a football game or a strike. Gestures and actions are easy to recognize and considered simple, whereas behavior and interactions are intermediate. Multi-person activities such as human–human interaction, group activities, or events are highly complex [1]. Considering the above-mentioned subactivities, the approaches used to recognize these vary widely. Such as basic methods include feature-based image processing techniques, background/foreground subtraction, action detection, and classification (i.e., optical flow, spatiotemporal interest points) [2–4].

✉ Usama Ijaz Bajwa
usamabajwa@cuilahore.edu.pk

Gulshan Saleem
gulshnsaleem26@gmail.com

Rana Hammad Raza
hammad@pnec.nust.edu.pk

1 Department of Computer Science, COMSATS University Islamabad, Lahore Campus, 1.5 KM Defence Road Off Raiwind Road, Lahore, Pakistan

2 Electronics and Power Engineering Department, Pakistan Navy Engineering College (PNEC), National University of Sciences and Technology (NUST), Habib Ibrahim Rehmatullah Road, Karachi, Pakistan

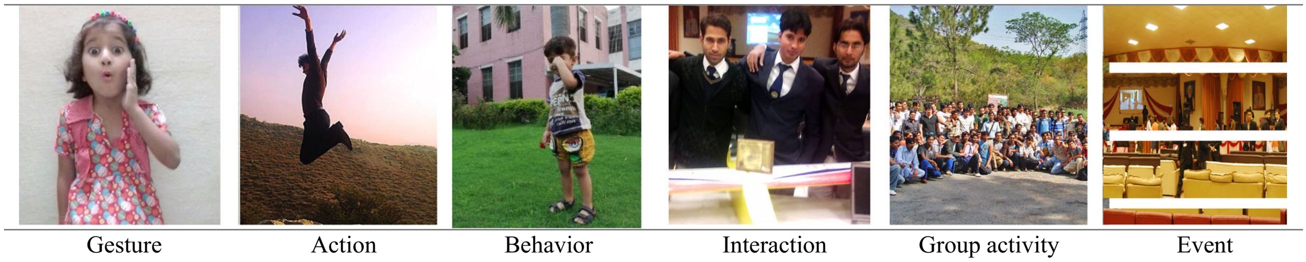

Fig. 1 Human activity recognition (simple to complex activities)

Advanced methods are a combination of multiple steps, which can collectively extract advanced features and perform in-depth analysis to recognize human activities [5–8]. Basic computer vision-based methods such as optical flow [9–11], spatiotemporal interest points (STIP) [12], hidden Markov model (HMM) [13], and advanced deep learning tools, for example, convolutional neural networks (CNN), recurrent neural networks (RNN) [14–16], are used to recognize human activity.

HAR has a multidisciplinary nature, and various daily life systems are influenced by performing HAR. HAR plays its role in indoor/outdoor environments, robotics, content-based information retrieval, human–computer interactions (HCI), security surveillance, video surveillance, educational sector, monitoring, and social interaction-based applications [17]. Hence, because of rapid technological advancement of daily life systems, there is a need for an up-to-date survey to discuss the progress of HAR and also to highlight its challenges [1]. Considering previous surveys, HAR systems can be classified as online or offline based on the input data and processing strategy. Then, there are unimodal/multimodal approaches that use different modalities, such as video frames, audio cues, skeleton data, and depth data. Most of the previous surveys have discussed handcrafted approaches, and few recent surveys have incorporated learning-based approaches as well [18–20].

1.1 Methodology for survey of HAR approaches

This study attempts to provide a method-based classification of approaches through taxonomy. It also provides a comparative analysis of state-of-the-art methods presented since 2011 to present an overview of HAR domain. This study includes 46 state-of-the-art methods, which are based on topics such as "Human activity Recognition," "Human Action Recognition," "Online Activity Recognition," "Learning-based Human Activity Recognition," and "Handcrafted features-based Human Activity Recognition." The existing state-of-the-art surveys are also discussed to analyze the up-to-date findings. Figure 2 shows the process of selection of studies ((a) Selection Process) and parameters used for the analysis of these studies ((b) Analysis Process). As shown in Fig. 2, studies are selected from multiple databases and initial selection is made on the basis of relevant topics. Then, all studies published earlier than 2011 have not considered and search is performed again through analyzing title of studies. This process helps to remove duplicates and survey studies from the selected set, and as a result, 2500 studies are left while others have been discarded. This survey is based on reviewing video-based HAR approaches and benchmark datasets. Therefore, selected set is refined to get studies that have used video benchmark dataset for evaluation of their model. For this purpose, we have reviewed the abstract and sometimes experiment too in case abstract does not provide necessary details.

As a result, 46 studies are selected for state-of-the-art analysis which includes studies on feature-based models, deep learning-based models, online activity recognition model, and methods for multimodal HAR. We have analyzed all selected studies using various parameters, such as publication year, method type, data input, activity level, dataset size, and its performance on benchmark datasets. These parameters help in identifying which activities among simple, intermediate, and complex are frequently used in research. The size of the dataset is an important indicator to determine what types of datasets are more useful across selected studies. Several evaluation measures are used for human activity recognition, such as Average Precision, which is the most reported measure. However, accuracy, recall, f-measure, likelihood ratio, and area under curve (AUC) are also popular among studies.

The major contributions of this study are as follows:

- This study highlights various approaches and proposes a HAR taxonomy, and elements of taxonomy are discussed with the help of HAR methods. HAR approaches are mainly divided into handcrafted/feature-based HAR and learning-based HAR approaches which are further sub-divided up to four levels to cover simple feature extraction-based methods such as trajectories and space–time feature.

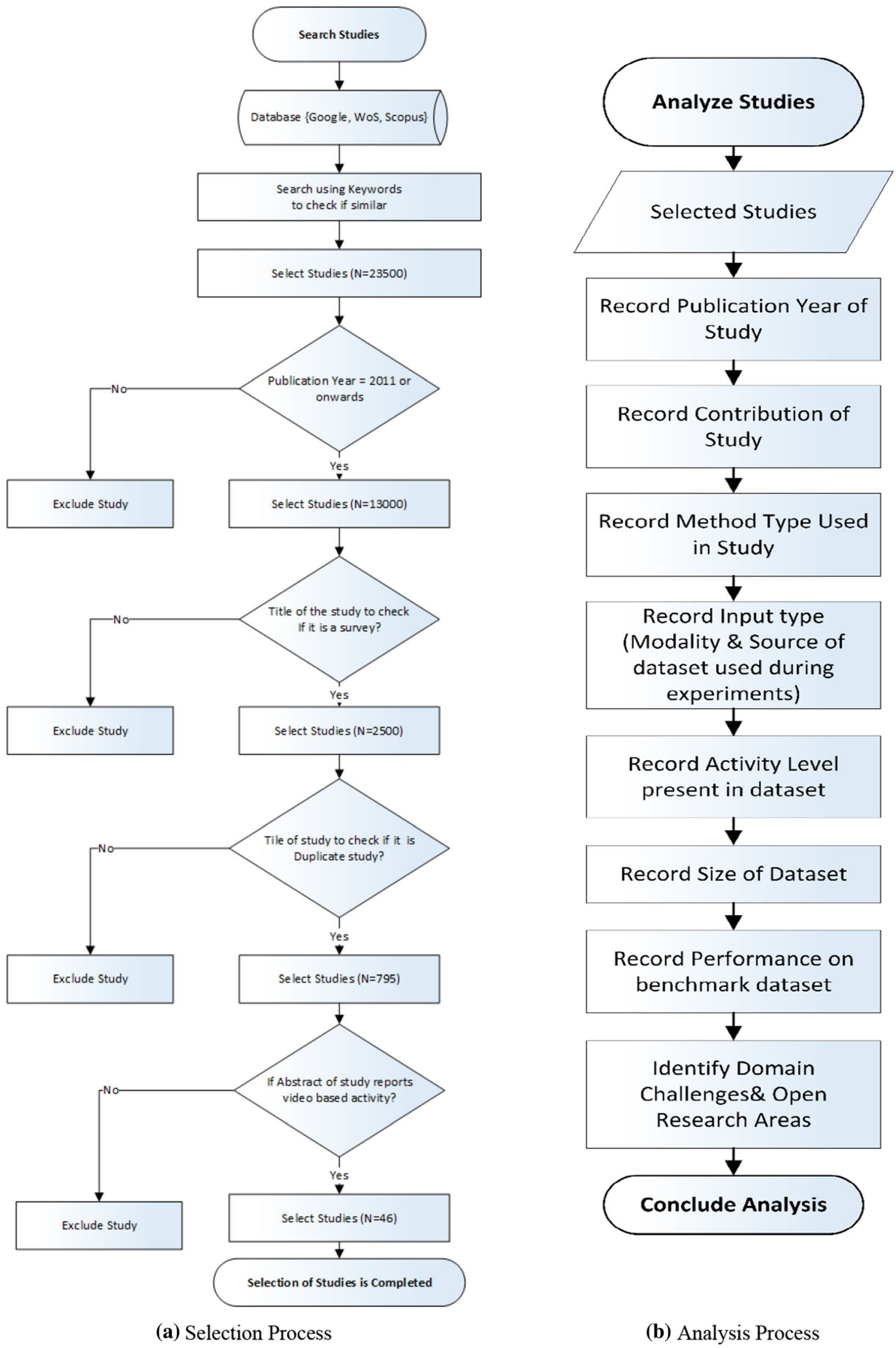

Fig. 2 HAR survey process for analyzing state-of-the-art methods

Table 1 State-of-the-art survey related to HAR

Author	Year	Activities	Complexity	Application	Contribution
Vishwakarma et al. [18]	2013	Abnormal Actions, Behavior, and interactions	Intermediate	Security Surveillance	Based on Activity recognition, object tracking, object detection tasks, and behavior understanding using handcrafted approaches. Few surveillance-based video datasets are also discussed
Ke et al. [21]	2013	Single person multi-person crowd activities (Actions, Interactions)	High	Pose estimation, Falling Detection, Security Surveillance	This survey provided details of Video-based activity and abnormal activity recognition methods
Vrigkas et al. [24]	2015	Actions Behavior	Intermediate	Action Recognition, Behavior Understanding	This survey categorized the HAR into unimodal and multimodal approaches and supports the effectiveness of later approach
Cheng et al. [22]	2015	Multi-type of activities including Actions, Interaction,	Simple	Action Recognition Systems	This survey focused on human action recognition-based approaches and few benchmark datasets have also been discussed
Zhu et al. [36]	2016	Actions	Simple	Action Recognition System	This survey covered the handcrafted and learned representations for human action recognition
Dawn and Shaikh [23]	2016	Actions	Simple	Action Recognition System	This survey discussed human action recognition with Spatiotemporal interest point (STIP) detector-based methods. Performance of selected methods has been discussed along with their results on different benchmarks
Sargano et al. [20]	2017	Actions, Interactions	Intermediate	Human activity Recognition	HAR approaches along with benchmarks have been discussed. Application areas have also been highlighted
Herath et al. [25]	2017	Multi type of activities including actions and Interaction	Intermediate	Daily Monitoring Systems, Activity Recognition Systems	This survey is focused on deep representation of action recognition domain. It provides the architectural details of different action recognition models along with performance on few benchmark datasets
Tripathi et al. [37]	2018	Abnormal Activities (Actions, Interaction, Group Activities)	High	Abandoned object Detection, Theft Detection, Violence Detection, Illegal Parking on Road Detection, Accidents Detection, Fire Detection	This survey is focused on suspicious activity recognition. Feature-based approaches along with classical machine learning methods have been described to explain state-of-the-art methods
Yao et al. [32]	2019	Daily activities Sports activities (Actions, Interaction)	Intermediate	Human Activity Recognition System, Daily activity monitoring system, Sports System	This survey provided Convolutional neural network-based action recognition along with performance of popular methods on large-scale datasets and highlighted the limitations and future directions
Moreno et al. [28]	2019	Daily activities (Actions, Interactions)	Intermediate	Human activity recognition system. Monitoring Systems	The survey has divided the approaches into three main categories, i.e., handcrafted features, depth sensors, and deep learning-based approaches which are further explained briefly

Table 1 (continued)

Author	Year	Activities	Complexity	Application	Contribution
Wang et al. [27]	2019	Abnormal Actions, Behavior, and interactions	High	Human behavior recognition	Focused on sensor-based behavior recognition and described the process of channel state-based behavior recognition. They categorized methods into model based, pattern based, and deep learning-based approaches
Liu et al. [29]	2019	Actions, gestures, and interactions	Intermediate	Daily activity recognition, Gesture recognition, User identification, Indoor localization & tracking	Focused on Wi-Fi signal processing-based activity recognition. Explained different setups of wireless sensing strategies such as RSSI-based, CSI-based, FMCW-based, and Doppler shift-based methods
Zhang et al. [33]	2019	Actions, Interactions, Group Activity	High	Human Activity Recognition System, Action Detection System	The survey discussed both action recognition and action detection, whereas action recognition is further extended toward action representation methods and interaction recognition methods
Jegham et al. [26]	2020	Multi activities (Actions, Interactions)	Intermediate	Human Activity Recognition System	Highlighted the constraint and challenges faced during the process of activity recognition. Action recognition approaches and few benchmarks have also been described
Dang et al. [30]	2020	Sensor-based data for Action Recognition, Multi Activities (Actions, Interaction)	Intermediate	Ambient Living Environment. Daily Monitoring System. Human Activity Recognition System	Based on sensor and vision-based HAR including benchmarks for both. Focused on feature Engineering and Preprocessing methods used for HAR
Beddiar et al. [1]	2020	Multi type of activities (Actions, Interaction, Group activity)	High	Human Activity Recognition	Provided general overview of HAR, including approaches, datasets, evaluation measures, and challenges of the domain
Das et al. [34]	2021	Actions, Interactions	Intermediate	Real-time human activity recognition. Daily activity monitoring	Focused on methods used for real-time human activity recognition. Presented challenges of real-time HAR
Chaurasia et al. [31]	2022	Multi type of activities (Actions, Interaction, Group activity)	Complex	Daily activities, Military activities, Abnormal activities, Ambution, Transportation activities	have worked on activity recognition and classification (ARC) smartphones and wearable sensors. Moreover, authors have concluded that ARC depends on the classification technique, number of sensors, device type, orientation, and placement. They have classified studies using ten parameters and highlighted domain challenges
Gupta et al. [35]	2022	Multi type of activities (Actions, Interaction, Group activity)	Complex	AI-based HAR applications, Hybrid AI models for HAR, Abnormal Human activities based	Authors have stated HAR design, dependability, and stability are major areas that need improvement to improve the HAR process

- This study also discusses HAR benchmark datasets, which have been used to perform experimentation and evaluation of methods. HAR datasets discuss their characteristics, e.g., single-view, multi-view, RGB, and RGB-D information, as well as instance-based details. Every dataset serves a purpose, and their brief description can help researchers to choose one accordingly.
- State-of-the-art methods are analyzed based on predefined parameters to highlight strength and limitations of domain. This survey includes 46 state-of-the-art approaches presented since 2011, and we have divided the methods into three categories: online/offline, unimodal/multimodal, and handcrafted feature-based approach/learning-based approach. The selected studies are further classified based on the complexity of the activity (i.e., simple, intermediate, or complex), as well as the size of the dataset (i.e., small, medium, large). It also includes the recognition rate (Average Precision) of selected studies to highlight how studies perform as compared to each other's. Hence, comparative analysis of various studies provides recent trends among the HAR research community and highlights open challenges for future research.
- The selected methods are classified as online/offline, unimodal/multimodal, and handcrafted feature-based approach/learning-based approach. The selected studies are further categorized based on activity complexity (i.e., simple, intermediate, complex) and size of dataset (i.e., small, medium, large). It also includes recognition rate (average precision) of selected studies to highlight their performance as compared to each other. Reported recognition rate may contribute toward significance of a selected study, but it is not a basis for comparison.

  We aim to provide the recent trend among HAR research community so that open challenges can be highlighted for future research.
- This study discusses HAR issues that were brought to light through comparison analysis, and it includes the environmental complexity of high intra-class variations and the inter-class similarity problem. Similarly, background, multi-view, and illumination variations are the primary issues that can affect the performance of the recognition system.

Section 2 provides a review of previous surveys and emphasizes the importance of this study. The characteristics of widely used video benchmarks for HAR are covered in Sect. 3. Section 4 then provides a taxonomy and detailed review of state-of-the-art HAR approaches to highlight research trends in HAR. Section 5 discusses the limitations of HAR and open research areas, and Sect. 6 concludes the study.

2 State-of-the-art HAR surveys

Human activity recognition is complex and involves variety of tasks. For example, action representation-based approaches need feature extraction and descriptors-based methods. Human activity analysis is complex and performed by using both machine learning and deep learning approaches, whereas we have conducted a survey on different approaches of HAR and categorized HAR into input processing strategy-based, modality-based, and model-based approaches. In previous years, authors have contributed toward HAR and presented specific to general

Fig. 3 RGB & RGB-D image from northwestern-UCLA [41]

Table 2 Characteristics of HAR benchmarks

Activity dataset	No. of Videos (Resolution)/FPS	No. of Actions (Actors)	View	Depth (D)	Activity types	Application Areas
KTH [42]	600 (160 × 120)/25	6 (25)	S	RGB	Actions	Human action recognition in outdoor conditions
Weizmann [43]	90 (180 × 144)/50	10 (9)	S	RGB	Actions	Human action recognition
UCF Sports [45]	150 (720 × 480)/10	10	S	RGB	Actions, Interactions (human–object)	Sports actions recognition
Olympic Sports [48]	783	16	S	RGB	Actions, Interactions (human–object)	Sports actions recognition
Hollywood [49]	233 (400 × 300, 300 × 200)/24	8	S	RGB	Actions, Behavior, Interactions, Group Activity	Activity recognition, Behavior Understanding, Interaction Recognition, Event Detection
UCF50 [50]	6681 (320 × 240)/25	50	S	RGB	Actions, Interactions (human–object)	Human Sports activity recognition
UCF101 [45]	13,320 (320 × 240)/25	101	S	RGB	Actions, Behavior, Interactions, Group Activity	Human activity recognition
YouTube Sports 1 M [51]	1,133,158	487	S	RGB	Actions, Interactions (human–object)	Human Sports activity recognition
IXMAS [47]	1650 (390 × 291)/23	13 (11)	M	RGB	Actions	Multi-view-invariant action recognitions
ActivityNet [52]	27,801 (1280 × 720)/30	203	S	RGB	Actions, Behavior, Interactions, Group Activity	Human activity and behavior understanding
YouTube 8 M [53]	~ 800,000	4716	S	RGB	Actions, Behavior, Interactions, Group Activity	Human activity and behavior understanding
HMDB51 [54]	6766 (320 × 240)/30	51	S	RGB	Actions, Behavior, Interactions, Group Activity	Human activity and behavior understanding
CASIA Action [55]	1446 (320 × 240)/25	8 (24)	M	RGB	Actions, Behavior, Interaction	Human behavior and interaction-based systems
AVA [56]	430	80	M	RGB	Actions, Interactions	Poses, person to person interaction and person-object interaction Recognition
UCF Crime [57]	1900	13	S	RGB-	Actions, Behavior, Interactions, Group Activity	Security Surveillance
UTKinect [44]	200 (320 × 240)/30	10 (10)	S	RGB-D	Actions	Human actions
MSR Action 3D [58]	567 (640 × 480)/15	20 (7)	S	RGB-D	Actions	Sports Gesture recognition
MSR Action Pairs [59]	180 (320 × 240)/30	10 (12)	S	RGB-D	Actions	Action pairs recognitions
SYSU- 3D HOI [60]	480 (640 × 480)/30	40 (12)	S	RGB-D	Actions, Interactions (human–object)	Daily activity Recognition
CAD-60 [61]	60 (640 × 480)/25	12 (4)	S	RGB-D	Actions	Daily activity recognition
CAD-120 [62]	120 (640 × 480)/25	10 (4)	S	RGB-D	Actions	Action labeling, human and object tracking
UTD-MHAD [63]	861 (512 × 424)/30	27 (8)	S	RGB-D	Actions, Interactions	View- invariant human action recognition
RGB-D HuDaAct [64]	1189 (640 × 480)/30	12 (30)	M	RGB-D	Actions, Interactions	Daily activity recognition

Table 2 (continued)

Activity dataset	No. of Videos (Resolution)/FPS	No. of Actions (Actors)	View	Depth (D)	Activity types	Application Areas
Berkeley MHAD [65]	660 (640 × 480)/30	11 (12)	M	RGB-D	Behavior	Human behavior Recognition
Northwestern-UCLA [41]	1475 (640 × 480)/30	10 (10)	M	RGB-D	Actions, interactions	Cross- view action recognition
UWA3D Multi-view [46]	900 (640 × 480)/30	30 (10)	M	RGB-D	Actions	Similar and cross-view action recognition
LIRIS [66]	9800 (640 × 480, 720 × 576)/25	828 (21)	M	RGB-D	Actions, Interactions	Human activity recognition
G3Di [67]	574 (640 × 480)/30	12 (15)	S	RGB-D	Actions, Interactions	Gaming interaction activity
NTU RGB + D [68]	56,880 (512 × 424, 1920 × 1080)/30	60 (40)	M	RGB-D	Actions, Behavior, Interaction	Daily Activity Recognition, Health surveillance systems
ShakeFive [69]	100	2 (37)	S	RGB-D	Actions	Handshake Recognition

Fig. 4 HAR datasets categorization

survey-based studies, which are discussed in this section, and Table 1 summarizes these surveys.

Action representation-based survey

In 2013, Vishwakarma et al. [18] have published a survey on surveillance-based activity recognition that primarily covers classical HAR approaches. They have classified HAR approaches as hierarchical or non-hierarchical. It provides a review of motion detection and object tracking methods, and characteristics of a few HAR datasets have been discussed. Ke et al. [21] published a survey to provide a general framework of HAR, which includes object segmentation techniques, feature extraction techniques, activity detection techniques, and classification techniques. Authors have thoroughly discussed the handcrafted approaches used in HAR in both [18] and [21] surveys. Cheng et al. [22] have discussed similar approach as used in [18] and provided characteristics of action recognition benchmarks. Dawn and Shaikh [23] used spatiotemporal interest points (STIP) to emphasize the effectiveness of STIP detectors as STIP detectors can improve HAR tasks because of their robustness.

Handcrafted vs. learned representation-based Survey

Vrigkas et al. [24] presented their findings based on unimodal and multimodal approaches that are further subdivided to discuss HAR. The survey's focus is skewed toward multimodal approaches because they provide a better feature set for learning. They have highlighted challenges faced by multimodal approaches, such as computational cost. It includes both traditional ML and advanced deep learning models, i.e., CNN. In addition, the survey provides a review of a few publicly available datasets that can be used for HAR. Zhen et al. [19] published a survey that has discussed two major HAR approaches: learned representation and handcrafted representations. Each one is further subdivided to analyze both categories and highlighted the strength of deep learning-based approaches. The survey in [19] was the first survey to compare traditional approaches with modern deep learning-based approaches. Similarly, Sargano et al. [20]

Fig. 5 UCF-101: single-view dataset [45]

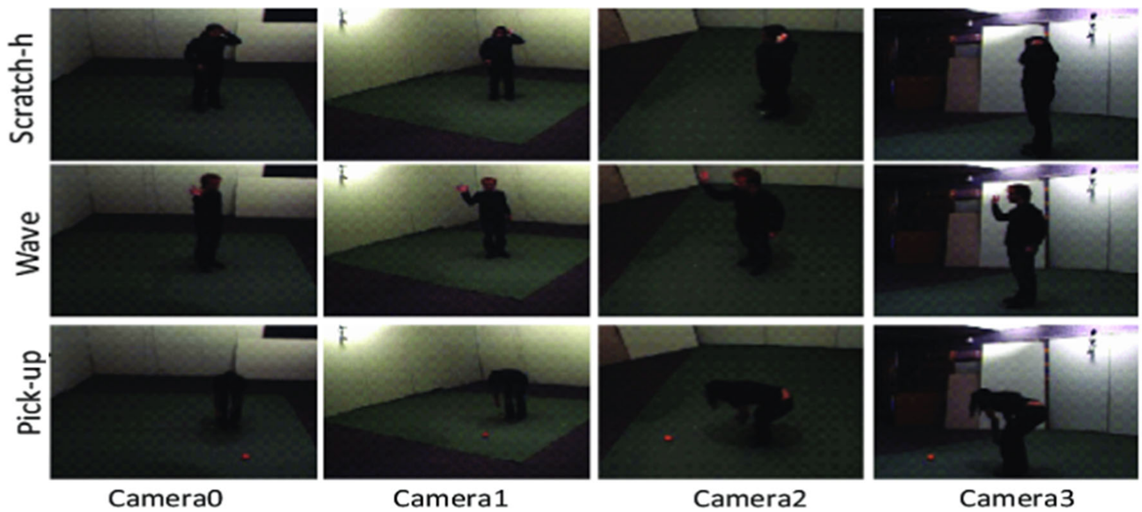

Fig. 6 IXMAS: a multi-view dataset [47]

presented a survey on handcrafted vs. learning-based approaches in 2017. They have discussed few publicly available HAR datasets and popular HAR applications. In contrast to [20], Herath et al. [25] also conducted a survey focusing on deep representation of action recognition. It has thoroughly discussed the popular handcrafted HAR features as optic flow, motion history image, trajectories, and other motion descriptors. They have also shown the architectural differences between popular networks like spatiotemporal networks, multiple stream networks, deep generative networks, and temporal coherency networks. Then, in 2020, Jegham et al. [26] have attempted to provide a quantitative analysis of a few popular methods while also discussing their applicability in various scenarios. The primary goal of their work is to highlight HAR issues through comparative analysis.

Sensor-based survey

Authors in [27] have surveyed channel state-based behavior recognition and thoroughly described the concept of channel state information. They have provided details of methods used for channel state-based behavior recognition and categorized it as refined behavior recognition, coarse behavior recognition, and inference activity. They have described channel state information-based behavior recognition with the help of three application areas which are model based, pattern based, and deep learning-based approaches. Authors have considered five major aspects for describing behavior recognition application, which are experimental equipment, experimental environment, behavior type, classifier, and performance. The authors in [28] have discussed sensor-based HAR systems and showed handcrafted feature-based approaches and deep learning-based approaches. Authors in [29] have presented a survey on HAR through wireless signal (e.g., Wi-Fi) as motion of the human body affects the wireless signal propagation. The authors have described the basic strategy and structure of wireless sensing environment for HAR. They have presented a variety of HAR applications which can be recognized by using wireless sensing technology such as fraud detection, daily activity monitoring. The authors have categorized sensing strategies based on HAR into received signal strength indicator-based (RSSI), channel state information-based (CSI), frequency shift for

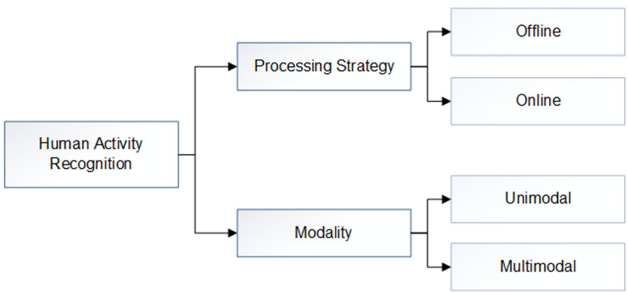

Fig. 7 Input data variations within HAR approaches (baseline methods are shown in Fig. 8)

frequency-modulated carrier wave-based (FMCW), and Doppler shift-based method. They have also added recent HAR methods based on these sensing variations and highlighted limitations of Wi-Fi-based approaches. Dang et al. [30] have discussed both vision- and sensor-based HAR systems along with corresponding HAR approaches and datasets. Chaurasia et al. [31] have worked on activity recognition and classification (ARC) smartphones and wearable sensors which include basics of ARC along with wearable and inertial sensors of smartphones. Moreover, authors have concluded that ARC depends on classification technique, number of sensors, device type, orientation, and placement.

Other

This category includes survey which covers multiple areas within HAR or is hard to classify among above mentioned categories. Authors in [32] discuss convolutional neural network-based HAR methods and their performance on large-scale datasets along with their performance on large-scale datasets. Zhang et al. [33] have investigated action classification and detection methods using RGB and depth-based datasets. Authors have discussed both handcrafted and deep learning methods for action classification and detection. The authors have shown the importance of action detection strategies for improving HAR performance. Beddiar et al. [1] have summarized HAR by discussing methods and benchmark datasets. They have identified HAR's limitations and challenges, which can be explored to extend research. The authors have focused on HAR process and presented methods for action detection and classification. Authors in [34] have presented HAR survey to highlight the recent trends for real-time human activity recognition. They have described various types of methods and evaluated their application to real-time scenarios. Their survey also highlighted the challenges of real-time online HAR, such as processing time of a method. Gupta et al. [35] have worked on analyzing human activity recognition to highlight future directions and explained three major points which needs improvement. Authors have stated HAR design, dependability, and stability are major areas, which needs improvement to improve HAR process.

All above-discussed surveys are summarized in Table 1, which presents highlights of each survey. Based on the foregoing, it is possible to conclude that most of the surveys lack a general perspective of the domain and cannot combine all elements of the HAR system within a single

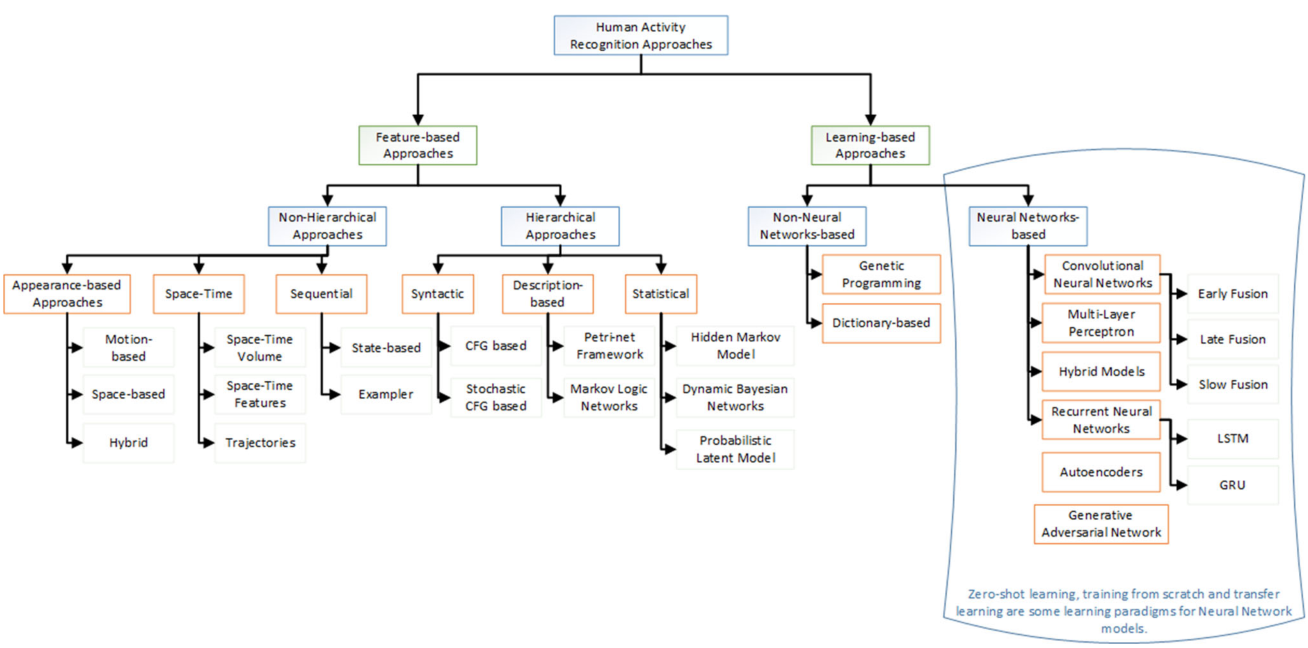

Fig. 8 Proposed taxonomy of human activity recognition approaches

study. As a result, this survey attempts to combine all necessary elements of HAR to show its multidisciplinary nature. These elements include feature-based methods, classification-based methods, multi-modality-based methods, online learning-based methods, dataset used for these methods, and state-of-the-art approaches of HAR. Moreover, it attempts to highlight limitations of HAR and provide open research directions.

3 Activity recognition datasets

So far, many benchmark datasets have been published, covering a wide range of activities. The choice of dataset influences the selection of a suitable approach for human activity recognition. Regarding dataset, the HAR presents several challenges, such as inter/intra class variations and the environmental setup used while recording actions (indoor/outdoor, camera, view angles). Inter-/intra-class variation occurs because of the unique nature of each human. For instance, when walking, some people take small steps while others take gigantic steps, some people avoid obstacles while others jump over them. More action classes may have overlapping, for example, complex actions comprise small actions. For example, fighting class involves punching, kicking, and thrusting. HAR datasets involve a lot of variations, which are explained in few previous surveys. In [38], authors divided datasets into three categories: heterogeneous actions and specific actions, and others. The heterogeneous actions include different types of actions, for example, walking, jumping running, etc. Specific actions include application-based datasets such as datasets of crowd behavior, abandoned objects, activities of daily living (ADL), fall detection, and pose & gesture, whereas other categories have datasets of motion capture (MOCAP), infrared and thermal. Authors have discussed HAR datasets and explains their characteristics. It includes publishing year, number of videos, actors involved, type of actions, application area, view information, and ground truth data of HAR datasets. Authors have presented a variety of methods used for each dataset. In [38], datasets were classified based on actions, whereas in [39], authors have discussed RGB-D (Fig. 3) video datasets. They have included characteristics of 27 single view action datasets, 10 multi-view datasets, and 7 multi-person datasets. It contains information about publishing year, number of videos, actions, and actors, and dataset complexity issues. It provides details of dataset splits (i.e., test, train, validation) and discussed some HAR methods for each dataset. In another survey [40], authors have classified datasets into RGB and RGB-D to discuss challenges of HAR datasets. They have highlighted five distinct challenges of datasets which are illumination, view variation, occlusion, annotations, and fusion of modalities. They have discussed HAR methods for each dataset and also discussed HAR studies to highlight dataset challenges. Considering available reviews on HAR datasets, this study highlights the major findings of datasets and provides discussion to support HAR benchmark analysis. Therefore, we have shown major classification of datasets in Table 2 using attributes such as image resolution, camera view, modality and type of activity, and the respective application areas, etc.

This survey presents a collection of HAR benchmark datasets organized by data view or data acquisition mode, i.e., single view or multi-view. Figure 4 illustrates the HAR task variation across datasets, including type of activities and modality variations. Few HAR datasets are discussed below under single view and multi view datasets.

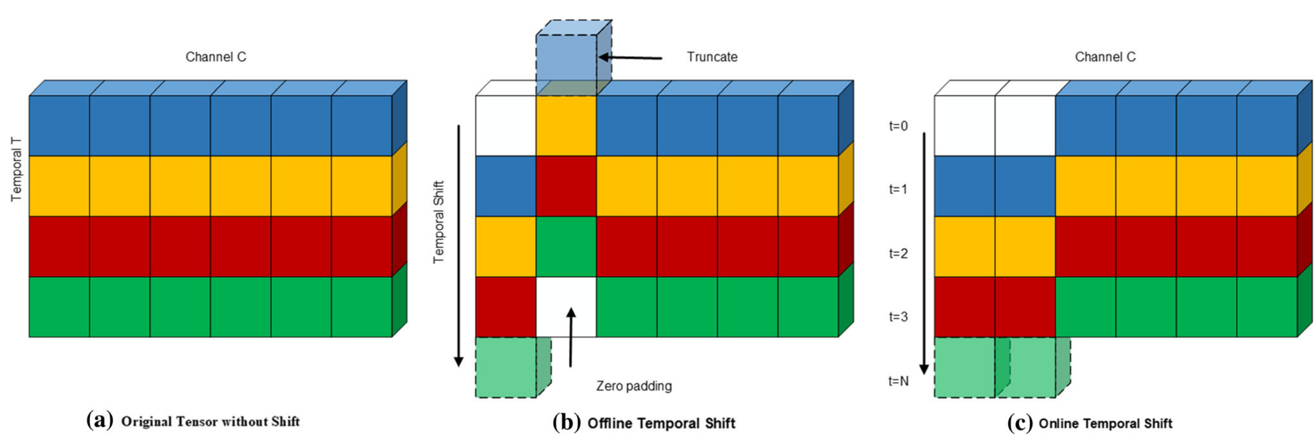

Fig. 9 Online vs offline HAR model [71]

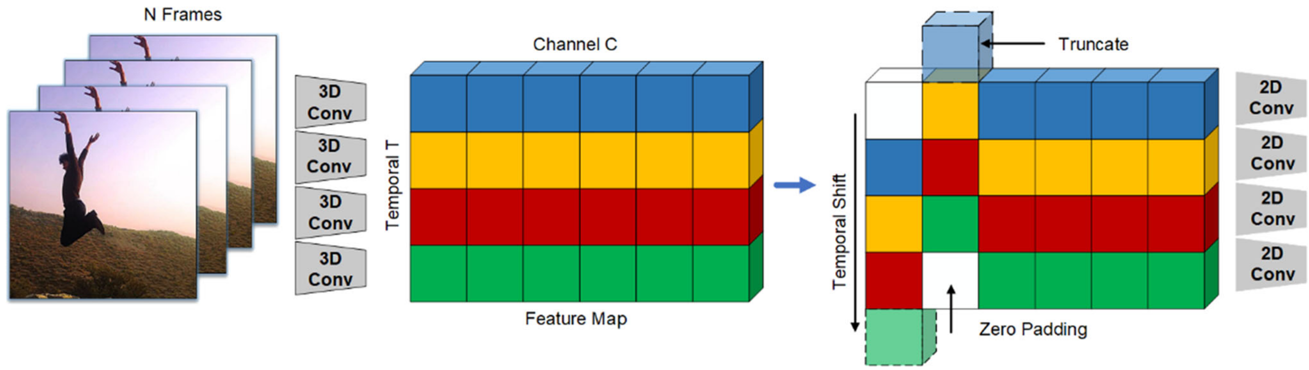

Fig. 10 Online human activity recognition framework [71]

3.1 Single-view action dataset

The single-view dataset is captured using a single camera and single view so does not involve view complexity in sequences, as shown in Fig. 4.

3.1.1 KTH

The KTH dataset [42] is based on six different actions (i.e., walking, running jogging, boxing, waving, and clapping), and these actions are performed by 25 actors. While performing these actions, four different background variations are used which are indoor, outdoor, scale variation within outdoor, and trying different clothes. The dataset includes 2391 videos captured through a static camera at a rate of 25 frames per second (fps) with a resolution of 160 × 120. The dataset is provided in training (performed by 8 persons), validation (performed by 8 persons), and test (performed by 9 persons) splits, but it does not include extracted silhouette of different actions and background of action.

3.1.2 Weizmann

The Weizmann dataset [43] includes 10 types of actions including running, walking, skipping, forward jump, up-down jump, galloping, 2-hands waving, 1-hand waving, and leaning performed by nine actors. The dataset includes total of 93 sequences captured through a static camera with 180 × 144 resolution of 25fps rate with an additional 10 sequences of walking (captured from different viewpoint). The background of the captured data is subtracted and the actions happening in the background are also included in the dataset.

3.1.3 UTKinect

The UTKinect dataset [44] is composed of 10 different actions, which include activities like walk, sit-down, stand-up, carry, throw, pull, push, wave and clapping. The actions were performed by 10 actors and each action is performed twice, which is performed through a variety of views and it includes 200 sequences. Along with action videos, labels and actions happening in the background are also included in the dataset.

3.1.4 UCF 101

The UCF 101 [45] dataset includes 13,320 RGB videos with 101 action categories which belong to 25 different groups and each group has 4–7 videos each. All actions belong to five major groups human–human interaction, human–object interaction, body motion, sports, and playing music as shown in Fig. 5. The UCF 101 provides realistic action videos rat6her than staged videos which improve the overall recognition task.

3.1.5 Multi-view datasets

The multi-view datasets are usually captured in one of two ways: multiple cameras at different angles or by using different viewpoints, as shown in Fig. 6.

3.1.6 UWA3D multiview

The UWA3D Multiview dataset [46] includes variety of sequences that were captured in a row with no pause. All these actions were performed by 10 different actors. They have performed different actions which include punching and waving with one hand, sitting down, and standing up, holding chest, walking, turning around, drinking, bending, running, holding head, holding back, kicking, jumping, moping floor, sneezing, sitting down (chair), squatting, two hands waving, two hand punching, vibrating, falling, irregular walking, lying down, phone answering, jumping jack, picking up, putting down, dancing, and coughing. This dataset is available in two versions: a single view version with 30 activities performed twice/thrice by actors,

Table 3 Quantitative analysis of state-of-the-art approaches of HAR

Reported paper	Year	Methodology	Online/Offline, Modality, HAR Approach, Activity Level, Training Data	Mean precision
Wang et al. [208]	2011	Dense trajectories are used to find actions from the data along with information of Histogram of oriented Gradient, optic flow, and motion boundary	Offline, Unimodal, Handcrafted features, Simple, Small	UCF Sports: 88.2%, Hollywood: 58.3%
Kliper-Gross et al. [209]	2012	The study is based on action recognition from unconstrained videos and representation-based architecture is used by extracting Motion interchange patterns from action data. Then General set of feature descriptors shows importance of feature set	Offline, Unimodal, Handcrafted Approach, Intermediate, Medium	HMDB-51: 29.2%, UCF-50: 68.5%
Oneata et al.[210]	2013	Performed action and event recognition through performing short action classification then locating these actions in lengthy movie videos along with recognition of complex events. It used Fisher vector instead of BoW and a set of handcrafted features is used to process the input data which include motion boundary histogram and SIFT. The data are normalized using L2-normalization method and classified through linear classifier. Also performed another approach by excluding human detectors	Offline, Unimodal, Handcrafted features, Complex, Medium	HMDB-51: 55.9% UCF-50: 90.5% Hollywood: 63% Olympic Sports: 91.2%
Wang and Schmid [154]	2013	Based on extraction of optic flow information which encodes the motion pixel value, and it is combined with extracted trajectories of data for action recognition (Trajectories, HoF feature descriptors)	Offline, Unimodal, Handcrafted features, Simple, Medium	HMDB-51: 57.2% UCF-50: 91.2%, Hollywood: 64.3% Olympic Sports: 91.1%
Jain et al. [211]	2013	Worked on extracting motion-based information using representation-based method to detect the actions from data. Finite set of feature descriptors are incorporated which includes HoG, Traj, MBH, HoF, and DCS. The extracted features are further fed to VLAD encoding technique	Offline, Unimodal, Handcrafted features, Simple, Medium	HMDB-51: 52.1%, Hollywood: 62.5%
Peng et al. [212]	2014	Performed action recognition by using representative-based method along with stacked Fisher vector (SFV) and Fisher Vector to extract action representations. SFV provides refined representation and abstract semantic information in layered manner to provide mid-level as well as high-level activity recognition	Offline, Unimodal, Handcrafted features, Complex, Medium	HMDB-51: 66.8%
Simonyan and Zisserman [213]	2014	Extracted appearance-based information from still frames and motion information of frames. Performed action recognition from videos by using deep neural network along with transfer learning. Authors have used two stream convolutional neural network to perform action recognition moreover multi-task learning is used to improve the results by adding classes from both action datasets i.e., HMDB-51, UCF-101	Offline, Unimodal, Deep learning, Intermediate, Medium	HMDB-51: 59.4%, UCF-101: 88.0%

Table 3 (continued)

Reported paper	Year	Methodology	Online/Offline, Modality, HAR Approach, Activity Level, Training Data	Mean precision
Karpathy et al.[51]	2014	Have used deep network (CNN) based on spatiotemporal information to perform action recognition from large-scale videos and opted for slow fusion-based learning strategy	Offline, Unimodal, Deep learning, Intermediate, Large	Clip Hit Sports-1 M: 41.9%, Sports-1 M: 60.9%, UCF-101: 63.3%
Sun et al. [214]	2015	Have worked on factorized spatiotemporal convolutional networks (FstCN) which perform factorization of original 3D Kernel into 2D Kernel for action recognition, i.e., Two stream clarifaiNet	Offline, Unimodal, Deep learning, Intermediate, Medium	HMDB-51: 59.1%, UCF-101: 88.1%
Wang et al. [215]	2015	Have worked with deep convolutional network to perform action recognition and used Two streams GoogleNet and two stream VGG-16. The aim is to overcome the overfitting problem of action recognition due to small size data so proposed to perform pretraining of both spatial and temporal nets using low learning rate and high drop-out ratio along with data augmentation	Offline, Unimodal, Deep learning, Intermediate, Medium	UCF-101: 91.4%
Wang et al. [196]	2015	Have proposed to use trajectory pooled deep convolutional descriptor (TDD) for action recognition and another method is used which implies the use of TDD along with histogram of optic flow to perform action recognition	Offline, Unimodal, Handcrafted features, Intermediate, Large	HMDB-51: 65.9%, UCF-101: 91.5% Conv pooling hit. Sports-1 M: 72.4%
Yue-Hei-Ng et al. [216]	2015	Worked on handling of full-length videos and proposed two different methods in which one is based on finding the best design of CNN through convolutional temporal feature pooling architecture. And the second approach is aimed at providing video in form of ordered sequence of video frames which is done by using RNN (LSTM) and is combined with the output of CNN	Offline, Unimodal, Deep learning, Complex, Large	Sports-1 M: 73.1% LSTM (image + opt flow) UCF-101: 88.6%
Fernando et al.[217]	2015	Proposed to used video wide temporal information to follow sequences using ranking machine which assigns ranks to produce action representation and this method is named as Rank pooling	Offline, Unimodal, Handcrafted features, Intermediate, Medium	HMDB-51: 63.7%, Hollywood: 73.7%
Donahue et al. [218]	2015	Proposed a recurrent convolutional architecture aimed at providing large-scale visual learning (LRCN) which performs temporal dynamics learning along with convolutional perceptual representation of actions within videos	Offline, Unimodal, Deep Learning, Complex, Medium	UCF-101: 82.9%
Wu et al.[80]	2015	Proposed multi-stream architecture which can perform multimodal feature extraction and so used CNN to extract multi features from videos. Then LSTM is used for the learning of long-term temporal variations in data. Both methods are fused to perform activity recognition	Offline, Multimodal, Deep learning, complex, Medium	UCF-101: 92.2% Columbia Consumer Videos: 84.9%

Table 3 (continued)

Reported paper	Year	Methodology	Online/Offline, Modality, HAR Approach, Activity Level, Training Data	Mean precision
Jiang et al. [219]	2015	Representation-based method has been proposed to extract motion related information from data using global and local referencing to overcome the camera movement problem within unconstrained videos	Offline, Unimodal, Handcrafted features, Intermediate, Medium	HMDB-51: 57.3%, UCF-101: 78.5%, Hollywood: 55.2% Olympic Sports: 80.6%
Lan et al. [220]	2015	Proposed Multi-Skip feature stacking (MIFS) stacks the extracted features in form of differential filters which helps in preventing data loss at coarse level. MIFS helps in action matching at different speed and ranges and speedup the process of feature extraction	Offline, Unimodal, Handcrafted features, Intermediate, Medium	HMDB-51: 65.1%, UCF-101: 89.1%, UCF-50: 94.4% Hollywood: 68%3, Olympic Sports: 91.4%
Tran et al. [221]	2015	Proposed deep three-dimensional convolutional neural network (3D ConvNet) for spatiotemporal feature extraction (C3D) which uses a linear classifier and perform significantly improved action recognition	Offline, Unimodal, Deep learning, Intermediate, Medium	UCF-101: 90.4%
Soomro et al. [72]	2016	Few frames are converted into super-pixel which are combined with spatiotemporal points to extract action segments and then dynamic programming on SVM score is performed for action prediction. Pose and appearance data are incorporated in online manner	Online, Unimodal, Handcrafted features, simple, Small	UCF-Sports 83.7%
Fernando and Gould [222]	2016	Proposed temporal pooling layer which can be incorporated with any convolutional neural network such as VGG-16 and AlexNet. The pooling layer is used to encode temporal semantics from long videos which are converted into fixed-length vectors	Offline, Unimodal, Deep learning, Intermediate, Small	UCF Sports: 87%, Hollywood: 40.6%
Fernando et al.[223]	2016	Proposed a hierarchical rank pooling that can extract the dynamics of CNN features using rank pooling function from video sequences. Then rank pooling is combined with non-linear feature function to provide video encoding mechanism	Offline, Unimodal, Deep learning, Complex, Medium	HMDB-51: 66.9%, UCF-101: 91.4%, Hollywood: 76.7%
Li et al.[224]	2016	Proposed a video representation framework VLAD based on linear dynamic system and helps in capturing video data using short medium and long ranges which includes motion and deep features of a video	Offline, Unimodal, Handcrafted feature, Complex, Large	Thoumas15: 80.8%, Olympic Sports: 96.6%, UCF 101: 90.9%
Feichtenhofer et al. [225]	2016	Worked on fusion methods of convolutional networks and proposed to use spatiotemporal network can be used at convolutional layer rather than SoftMax layer	Offline, Unimodal, Deep learning, Intermediate, Medium	HMDB-51: 69.2%, UCF-101: 93.5%
Varol et al.[226]	2017	Proposed LTC-CNN model for video representation based on long-term temporal convolutions (LTC), moreover raw pixels, optic flow estimation features are incorporated within model to improve action recognition	Offline, Unimodal, Deep learning, Simple, Medium	HMDB-51: 67.2%, UCF-101: 92.7%

Table 3 (continued)

Reported paper	Year	Methodology	Online/Offline, Modality, HAR Approach, Activity Level, Training Data	Mean precision
Jalal et al. [70]	2017	Presented spatiotemporal multi-fused features to perform online activity recognition which includes joint features, torso and key joint-based distant features, HoG, and few others	Online, Multimodal, Handcrafted features, Complex, Large	MSR Actions 3D: 93.3%, 1 M-MSR Daily Depth Activity: 74.3%
Singh et al. [227]	2017	Proposed a novel graphical representation to perform abnormal activity recognition by introducing geometric structure along with motion and appearance-based information, whereas activity classification is performed through SVM and global abnormal activity through Bag of Words (BoG) using STIP, SIFT, and DT feature set	Offline, Unimodal, Handcrafted features, Intermediate, Small	UCSD ped1: 97.14%, UCSD ped2: 90.13%, UMN: 95.24%
Carmona et al. [228]	2018	Have worked on improved dense trajectories (IDT) through incorporating more Temporal Templates-based features and three templates are constructed in form of third order tensor	Offline, Unimodal, Handcrafted features, Intermediate, Medium	KTH: 97.5%, Weizmann: 98.8%, HMDB-51: 65.3%, UCF-101: 89.3%
Zolfaghari et al. [74]	2018	Have proposed online recognition architecture (ECO) which uses feature representation from all video frames which are then fed to CNN network. The model uses half frames from the current sequence and half from the incoming queue data to reduce the overhead	Online, Unimodal, Deep learning, Intermediate, Intermediate	UCF-101: 93.3%, HMDB-51: 68.7%
Mukherjee et al. [81]	2018	Has proposed a motion capture strategy and produced dynamic images from RGB and depth videos separately using ResNet 101 network. The dynamic image reduces complexity by extracting sparse matrix from video, and resultant framework is fast and memory efficient	Offline, Multimodal, Deep learning, Complex, Medium	MSR Action 3D: 96.17%
Zhang et al. [82]	2018	Proposed a semantic-based multistream deep neural network for action attribute learning and action recognition along with zero shot action recognition. It also combines semantics in graph regularization and joint learning is achieved by using ADMM optimization algorithm	Offline, Multimodal, Deep learning, complex, Medium	MRA: 72.03%, UTA: 81.89%, MRP: 94.69% Accuracy: MSR Actions 3D: 93.40%, UTA Action 3D: 87.88%, MSR Action Pairs: 99.44%
Mao et al. [229]	2018	Proposed a deep convolutional graph neural Network and used self-attention graph pooling mechanism for action classification	Offline, Unimodal, Deep learning, Complex, Large	Youtube-8 M: 87.7%
Siddiq et al. [230]	2019	Proposed a feature selection approach named normalized mutual information-based feature selection (NMIFS) which is extended form of both max-relevancy and min-redundancy. Combination of Curvelet transform, LDA, and HMM is used to prove the state-of-the-art	Offline, Unimodal, Handcrafted features, Simple, Small	Accuracy KTH: 99%, Weizmann: 98.2%
Lin[71]	2019	Have proposed temporal shift module (TSM) to achieve efficiency with high performance. TSM provides 3D CNN performance, but it costs 2D CNN which are further categorized as unidirectional TSM (Online Recognition) and Bi-directional TSM (offline recognition)	Online, Multimodal, Deep learning, Complex, Large	Accuracy: Something-Something V2: 50.7%, Kinetics: 76.3%

Table 3 (continued)

Reported paper	Year	Methodology	Online/Offline, Modality, HAR Approach, Activity Level, Training Data	Mean precision
Franco [83]	2020	A multimodal approach is based on use of two stream data i.e., skeleton data and RGB data. Skeleton data provide human posture-based data, whereas RGB provides temporal information for the evaluation of action hence improve the action recognition	Offline, Multimodal, Handcrafted features, Complex, Small	CAD-60: 98.8%, CAD-120: 85.4%, Office Activity: 90.6%
Zhang et al. [231]	2020	Based on improvement in bad sample problem arises due to random cropping technique and for that motion patch-based Siamese convolutional neural network (MSCNN) has been proposed. Motion patch uses the idea of extraction of critical motion square region	Offline, Unimodal, Deep learning, Simple, Medium	Pretrained on Kinect UCF-101: 96.8%, HMDB-51: 74.8%
Arzani et al.[232]	2020	Worked on human–robot interaction system to handle both simple and complex activities and used probabilistic graphical models (PGMs) to design a structured prediction strategy. A deterministic switch is used to identify simple and complex activity subspaces considering all possible activities	Offline, Unimodal, Handcrafted features, Complex, Small	UT Kinect: 100%, Florence 3D Dataset: 96.11%, CAD-60: 97.6%
Gowda et al. [233]	2020	Have proposed a model SMART which provides efficient frame selection strategy from videos and it is based on temporal segment network (TSN and Kinetics)	Offline, Unimodal, Deep learning, Complex, Large	Accuracy: ActivityNet: 84.4%, UCF-101: 98.6%, HMDB-51: 84.36
Gowda et al. [198]	2021	Worked on zero shot learning using reinforcement method and proposed a clustering framework (CLASTER) which can take all training data at once rather than using individual optimization. They have trained their model on activity recognition benchmark datasets and then tested on unseen examples from real world which have made it complex but close to real-world scenarios	Offline, Unimodal, Deep learning, Complex, Medium	Accuracy When Tested on unseen data while Training data: [Olympics Sports: 68.8%, HMDB-51: 53.3.4%, UCF-101: 69.3%]
Wharton et al. [234]	2021	Have proposed a Coarse Temporal Attention Network (CTA-Net) which is aimed at capturing high level temporal data to learn useful spatial and temporal variations in a video	Offline, Unimodal, Deep learning, Complex, Large	SBU Kinect Interaction: 92.9%
Ullat et al. [235]	2021	Have proposed sequential extraction method which uses optical CNN model and Deep Skip Gated Recurrent Unit is proposed to perform sequential pattern learning	Online, Unimodal, Deep learning, Complex, Large	HMDB-51: 64.98%, UCF-101: 86.39%, UCF-50: 91.29%, Hollywood2: 68.21%, YouTube Actions: 92.63%
Khan et al. [236]	2021	Have worked on feature extraction process to improve action recognition and used shape features along with deep learning features to improve learning. Such as entropy controlled LSVM maximization is used for robust feature extraction	Offline, Unimodal, Handcrafted features, Intermediate, Medium	KTH: 98.66%, Weizmann: 99.1%, UCF Sports: 99.12%, UT Interaction: 100%
Ullah et al. [237]	2021	Have proposed a multi-view action recognition method which performs frame level feature extraction to feed these forward to conflux LSTM. Then correlation coeeficient is computed using view inter-reliant pattern learning and then action classification is performed	Offline, Multimodal, Deep Learning, Intermediate, Medium	MCAD: 86.9%, Northwestern-UCLA: 88.9%

Table 3 (continued)

Reported paper	Year	Methodology	Online/Offline, Modality, HAR Approach, Activity Level, Training Data	Mean precision
Reinolds et al. [238]	2022	Authors have compared the performance of both video-based and audio-based activity recognition. They have performed classification process for both types of input by extracting features for each	Online, Multimodal, Deep learning, Intermediate, Large	Real-Life Violence situations: 89%
Siddiqi et al. [239]	2022	Authors have used mutual information algorithm and expanded max-relevance and min-redundancy methods to select optimal features. Features are extracted through symlet wavelet transform and later action classification is performed through hidden Markov model	Offline Unimodal, Hand crafted features, Simple. Medium	Kinect depth dataset: 98.2%
Khare et al. [240]	2022	Have proposed a multiresolution video analysis scheme and used local binary pattern (LBP) along Zernike moment (ZM)	Offline, Unimodal, Hand crafted features, Intermediate, Medium	KTH dataset: 96.38%, CASIA dataset: 98.82%
Deotale et al. [241]	2022	Have proposed a four step activity recognition method which involves frames conversion, human body detection, action recognition and then occurrence time of action using two stream data (i.e., RGB image and optic flow) through CNN-based network	Offline, Multimodal, Deep learning, Complex, Large	ActivityNet: 39.37%
Zhang et al. [242]	2022	Have proposed ActionFormer which is an efficient method for timely action recognition in a single shot setting. It aggregates multiscale feature representation and local self-attention information which is forwarded to a decoder to perform action recognition	Offline, Unimodal, Deep learning, Complex, Large	ActivityNet: 53.5%, THOMUS: 65.6%

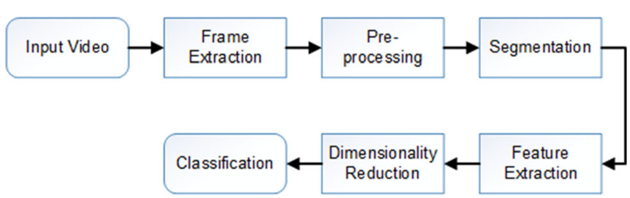

Fig. 11 General framework of handcrafted feature-based approaches

and a multi-view version with 30 activities performed by ten actors but captured four times using front, left, right, side, and top views to capture the sequences.

3.1.7 Northwestern-UCLA multiview action 3D

The Northwestern-UCLA Multiview Action 3D dataset [41] was designed by using three real-time cameras to capture 10 activities including pick up with one hand, pick up with two hands, drop trash, walk around, sit down, stand up, donning, doffing, throw, carry. The activities are performed by ten actors, and 1475 sequences are present in the dataset.

3.1.8 IXMAS

The INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset [47] includes 13 daily life activities which are checking watch, crossing arms, scratching head, sitting down, getting up, turning around, walking, waving, punching, kicking, pointing, picking, overhead throwing, and bottom up throwing. These activities are performed thrice by 11 actors and 2154 sequences are collected in the dataset. To capture the dataset, 5 synchronized and attuned fire wire cameras were used, and it also includes the silhouettes and visual hulls.

There are several benchmark datasets available that can validate the performance of human activity recognition methods, and a few of them are discussed in Table 2 to provide a short list of suitable datasets. Table 2 includes the number of videos and the number of classes, activity types, and modality information. Dataset characteristics may help in choosing a dataset while considering specific models such as large-scale datasets are appropriate for deep learning-based methods (e.g., CNN, RNN), whereas small-sized datasets are typically used to validate handcrafted feature-based approaches. Small datasets are ineffective for deep learning-based models, which require massive amounts of training data. The Weizmann dataset is a small dataset with 90 videos, whereas the YouTube 8 M dataset is the largest. Availability of large amounts of data is no longer a problem because of cheap CCTVs everywhere, but labeling that data remains difficult. As a result, the variety of datasets simplified the task and provided flexibility when validating any method. Along with the size of the dataset, the number of videos within a class is important when describing the quality of the dataset. It is preferable if each class within a dataset has an equal number of videos to avoid class imbalance.

3.2 Discussion

HAR benchmark datasets are complex to analyze as they try to mimic the real-life scenarios based on human activities. The purpose of HAR benchmark is to provide a close representation of human behavior in different scenarios. One of the most important aspects of a dataset can be its relation to reality, and a close relation of these two will provide a better human activity recognition. In daily life, illumination, scene variations, occlusion, and background activities vary widely. However, datasets may have not focused on such issues and were recorded in a controlled environment. Majority of HAR datasets are actor based, which means it includes activities, which are performed by different actors. For example, few daily life activity datasets do not focus on occlusion and background activities such as KTH [42] dataset, UT Kinect [44], and Northwestern-UCLA [41] datasets have a static background. KTH [42] and Weizmann [43] are small size action datasets, and most methods achieve 100% accuracy in these datasets. The reason is both datasets have a clear background with no occlusion and simple actions, which can be 100% classified by most of recent HAR methods. That is why both datasets can be used as a good start but cannot be up scaled for complex HAR scenarios.

Few datasets, which have considered occlusion and background variations, are useful for gaming/sports systems, e.g., MSR Action 3D [58] and G3Di [67]. MSR Action 3D dataset has RGB and depth information, but both channels are recorded separately, which causes synchronization problem. UCF-101 [45] and HMDB-51 [54] are daily activity-based datasets of intermediate size, which offer dynamic background and can be used for evaluating daily activity monitoring-based systems. The activities include human-to-human and human-to-object interactions and are useful for evaluating human computer interaction (HCI) systems. CAD-60 [61] is RGBD dataset of daily actions which are recorded in five different scene variations, but it has class imbalance problem.

NTU-RGBD [68] datasets is a large dataset with 56,880 videos recorded in a laboratory with strict guidelines, which made it partially useful for real-time activity recognition. It has daily life activities and health related actions such as falling and sneezing. NTU-RGBD [68] dataset can be used for evaluation of healthcare surveillance and daily activity monitoring systems. Sports 1-M [51] and YouTube 8 M [53] are large-scale datasets that offer background variation; occlusion and complexity of these datasets can be upscaled. Sports 1-M dataset has a substantial variation in sports action, which are annotated. The annotation or labeling is performed by content-based retrieval strategy, and therefore, it may be inaccurate. UCF Crime [57] is a large-scale dataset with 1900 videos of 13 different anomalies. The dataset offers inter class and intra-class problem, which may result in increased false positive rate. As UCF Crime has unbalanced dataset, which means few classes have significantly large amount of data as compared to others. Considering the above list of datasets mentioned in Table 2, there is a lack of 3D datasets captured in unconstraint environment. Majority of datasets avoid background and distant activities that are useful in real time scenarios, for example, surveillance systems.

4 Human activity recognition approaches

HAR is used in various daily life systems and can be performed by a variety of methods. It emphasizes the need for HAR taxonomy to discuss existing approaches. Previous surveys are focused on some specific tasks; for instance, [24] have discussed only unimodal and multimodal HAR approaches, [20] has used handcrafted vs learned representation to discuss HAR, [18] has used only single-layered & hierarchical approach-based division, and [1] has discussed both handcrafted vs learned representations and unimodal vs multimodal approaches. This study proposed a top-down taxonomy that can encompass all methods, from simple to complex. Figure 7 depicts input data variations within HAR, while Fig. 8 depicts taxonomy.

Human activity recognition can be done offline (via stored videos) or online (via a live stream), which is critical when dealing with real-time systems. Another variant is the source of modalities, which refers to either unimodal or multimodal methods. Unimodal methods rely on a single modality for input, whereas multimodal methods may use multiple modality inputs, for example, depth, audio cues, and skeleton data [24]. HAR includes simple offline-unimodal methods [51] as well as complex Online-multimodal systems [70] [71]. So, all HAR systems, whether Online/Offline or Unimodal/Multimodal, rely on handcrafted feature-based approaches or learning-based approaches. Baseline approaches used for above-mentioned systems are shown in Fig. 8. The taxonomy is divided into two categories: handcrafted feature-based approaches learning-based approaches. A unimodal or multimodal framework necessitates the careful selection of methods from handcrafted feature-based approaches or learning-based approaches. Both handcrafted and learning-based approaches are divided into sub-categories. Furthermore, recent learning-based methods, such as zero-shot learning and transfer learning, are significant. When all the activity classes are not available, such methods are useful. All above-mentioned methods are part of the HAR taxonomy to present a relationship between various variations. As it hasn't been done before, it has the potential to contribute significantly to the domain by demonstrating the multi-disciplinary nature of HAR. HAR taxonomy is discussed under qualitative analysis section through existing HAR methods to provide a brief description of each.

4.1 Online/offline processing strategy-based human activity recognition

Online human activity recognition uses the live stream which is fed to HAR model to perform activity recognition such as in augmented reality/virtual reality (AR/VR) and self-driving cars. Most of the methods are targeted to offline systems that process all video frames together and are not suitable for real-time systems, i.e., security surveillance. Soomro et al. [72] have used batch of frames from videos to estimate pose. They have used current frame to convert it into super-pixels along with conditional random fields to produce nodes and spatiotemporal points are used to extract actions. Short duration clips are used to predict action confidence via dynamic programming based on SVM scores. This approach has helped in capturing the sequential information of video, whereas appearance-based information and pose estimation are done online and only few frames are used for this purpose. Singh et al. [73] have addressed slow execution of offline approaches in real-time scenarios through multiple spatiotemporal action localization. To overcome these issues, CNN is used along with a single-shot multibox detector, which helped in construction and labeling of action tubes, which achieved real-time action recognition performance ranging up to 40 fps. Jalal et al. [70] have used Depth Differential Silhouettes (DDS) along with human temporal points to perform online activity recognition. It further considered the skeleton joint features, which include torso and key joint-based distant features. They have reduced the size of feature set through code vector. HMM is trained on these code vectors to recognize human activity segments through forwarding spotting and depth map is used for online activity recognition, whereas Zolfaghari et al. [74] have focused on long-term content along with fast video processing to perform efficient online recognition. They have proposed a 3D and 2D Combination Architecture (ECO) in which 2D network ensures feature representations from still images, whereas complex information is extracted from 3D network. To reduce the complexity and data overhead issue, half of the frames are taken from the current sequence and half from the upcoming sequence (Queue) to make predictions. Xu et al. [75] performed online HAR that was based on using temporal context of each frame while performing action detection in parallel. They have proposed a Temporal Recurrent Network (TRN) which is based on RNN. It works by predicting actions from each frame while anticipating future actions, so the future actions combined with historical data may produce better predictions. Lin et al. [71] have proposed a temporal shift module for both online and offline recognition. The offline recognition is bidirectional, whereas online recognition is unidirectional as it considers only upcoming video frames, as shown in Fig. 9.

TSM-based online recognition model is shown in Fig. 10, which provides low latency and low memory consumption rate as compared to other methods. Their model provides average precision of 95.5% while trained on UCF-101 dataset. It performs well on offline activity recognition with zero latency rate and 95.8% average precision.

To improve online action recognition from untrimmed videos, Gao et al. [76] proposed a Weakly Supervised Online Action Detection (WOAD) framework. It uses temporal proposal generator (TPG) that works offline to generate frame level labels and an online action recognizer (OAR) that detects online actions. Offline recognition is less complex than online recognition because it is based on stored videos, making dealing with such data easier compared to online. In offline scenarios, a decision is made after analyzing the entire video, whereas in real-time scenarios, recognition is required immediately based on new frames. Because action recognition from videos is primarily performed on stored data, most of the methods discussed in this study are offline, whereas as shown in Table 3, online approaches for video-based activity recognition are gaining popularity.

4.2 Modality-based human activity recognition

Most of the methods are offline and unimodal because these methods involve less complex computational strategies and resources. Unimodal approaches recognize activity by utilizing data from a single modality, for example, visual representation learned from image sequences or still images. Unimodal approaches perform well when motion-based features are used as methods based on space–time, stochastic, and shape-based data. Besides the methods mentioned, rule-based approaches, which include CFGs and statistical models (HMM) have performed well [24].

The research community's attention is shifting to multimodal approaches based on data from two or more modalities. Ofli et al. [65, 68, 77, 84] uses a variety of modalities to describe an activity, including RGB data, depth data, audio cues, skeleton data, optic flow, motion capture, and temporal data. Multimodal approaches primarily use two or three different sources of information to recognize actions by performing feature fusion such as early and late fusion, which can be classified as affective methods, behavioral methods, or social networking-based methods. Chen et al. have used facial expression along with action recognition to design emotion recognition system [63]. Rigkas et al. [78] have worked on behavior recognition using a fully connected conditional random fields (CRFs) model which can recognize friendly, aggressive, and neutral behaviors. In [79], joint sparse regression-based method has been proposed which uses depth data as well body parts information to extract variety of features for action recognition. Wu et al. [80] proposed a deep learning-based multi-stream architecture that can extract multiple features from videos using CNN to perform multimodal feature extraction. The extracted feature data are fed into the Long-Short Term Memory (LSTM) model, which uses this information to learn long-term temporal variations in data and then combines it to perform human activity recognition. Jalal et al. [70] have fused data of different modalities which includes torso-based distant feature descriptors, key joint-based feature descriptors, motion features, shape-based features, and a few others. Mukherjee et al. [81] have proposed the use of dynamic images by extracting motion information from RGB images and depth images separately, which are then combined. The task is performed by using two streams of Resnet-101 network and resulted in reduced sparse matrix from videos. Zhang et al. [82] have also worked with different modalities and produced a semantics-based multi-stream deep neural network for action attribute learning and action recognition along with zero shot action recognition. It also combines semantics in graph regularization and joint learning to use adam (adaptive moment estimation) optimization algorithm. Franco et al. [83] used temporal and posture-based data for activity recognition, and a two-stream architecture based on skeleton and RGB data was proposed. Skeleton data provide human posture information, whereas RGB data provide temporal information for action evaluation, resulting in an improvement in the action recognition process.

4.3 Model-based human activity recognition

Model-based human activity recognition involves methods of feature extraction from action data and classification of these data in specified class. In this study, handcrafted feature-based and learning-based approaches are discussed to cover wide range of HAR methods.

4.3.1 Handcrafted feature-based approaches

The feature-based/handcrafted approaches use statistical or image processing techniques to calculate features. Figure 11 depicts the general framework of feature-based human activity recognition. These methods rely on manual feature extractions which include different statistical, temporal, and appearance-based features.

4.3.1.1 Non-hierarchical approaches Non-hierarchical or single-layered approaches use raw video data and are classified into two types (i.e., space–time and sequential approaches) based on how the temporal dimension is considered, which are further classified into relevant groups of methods. Such methods are used to recognize short and simple human actions (e.g., running, jumping, walking) and are normally evaluated on small datasets, for example, KTH [42] and Weizmann dataset [43]. Non-hierarchical approaches are based on data representation and matching, which is normally done using a suitable feature extraction strategy. The non-hierarchical approaches can be used in different sequential combinations to recognize more complex actions.

Space–time approaches The space–time approaches are based on the problem's spatiotemporal nature. Because time is a regular domain, features can be extracted from a 3D volume containing a 2D spatiotemporal sequence of images with another equal set of pixels in the third dimension (XYZ plane). This means that the video has a spatiotemporal volume with important information for action recognition, and as a result, many researchers have contributed by proposing significant matching-based algorithms to identify underneath motion patterns.

Space–time volumes The space–time volume approaches consider the entire volume as a template or simply a feature that is then matched with previously existing videos to perform action classification. It is done by using a matching algorithm such as Bobick and Davis's [84] identified motion pattern. Hu et al. [85] contributed by combining the motion history images (MHI) and appearance-based information, whereas appearance still relies on two features, i.e., foreground image and Roh et al. [86] extended motion pattern strategy using volumetric motion template to provide view-independent action recognition and shifted MHI from 2 to 3D. Histogram of Oriented Gradients (HOG) to get the magnitude and direction of edges and corners of a specific action. Another famous combination was to use both global and local features and here global features involve the contour coding of motion energy image (MEI), whereas local features simply provide a bounding box for an action that further uses the multi-SVM for classification of feature points [87]. Then, Kim et al. [88] used the concept of representing spatiotemporal features gained from different actions by producing accumulated motion images (AMI).AMI pixel values are used to produce a rank matrix. This task is based on computing the distance value of the rank matrix of two videos, i.e., candidate video and the target video. Another group of researchers [89] has designed a pose descriptor by using rectangular patches that were extracted over human silhouettes and named that descriptor as Histogram of Oriented Rectangles (HOR). A similar approach presented by Fang et al. [90] based on silhouettes have been proposed which aimed at the mapping of high dimensional silhouettes to the spatial motion of low dimensional points to get the information about inherent motion structures. After the pose descriptor, Ziaeefard et al. [91] has used skeleton-based data and designed a cumulative skeletonized image (CSI) regarding time. This skeleton-based image is used to create distance-based histogram to feed the information to SVM model for matching. The authors have also used the idea of similar and dissimilar actions while matching processes. Two types of CSI histograms were taken for similar and dissimilar actions. Wang et al. [92] were made using the notion of "bag-of-words framework" so taking the word as frame and document as videos to design semi-latent topic models (STM) which resulted in an efficient action recognition system with better accuracy as well, but the drawback was the limited number of latent topics. Another research has been presented by Guo et al. [93] stating that the action is the deformation of local shape features (i.e., centroid-centered object silhouettes) over a temporal sequence. The feature set of 13-dimensional normalized geometric vectors is used to produce a covariance matrix that holds the shape of the silhouette. The Riemannian matrix is calculated between the

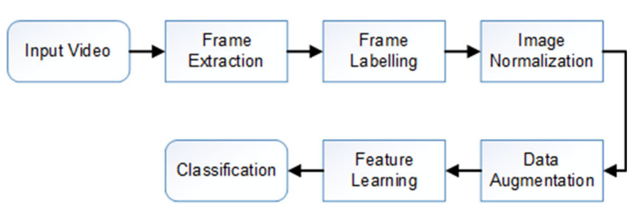

Fig. 12 General framework of learning-based approaches

covariance matrixes of two actions for classification. Another direction toward the action recognition was to improve the process of video analysis and for that purpose, Kim et al. [94] have proposed a method to check the similarity between two videos by using the assumption that similar videos represent similar actions through extending canonical correlation analysis. The idea was good as it helped in ignoring the irrelevant complexity within a task by avoiding explicit motion estimation within a frame.

Space–time trajectories The trajectory-based approaches use raw data from videos, and for that purpose, tracking points are obtained by considering joint positions of the human body. The tracking of such joints or interest points results in the construction of a trajectory. Similarly, Messing et al. [95] used KLT tracker to track Harris3D joint areas (feature trajectories) which produced log-polar velocities as sequence. Then learning of these velocities (i.e., velocity-history language) is performed by applying a generative mixture model to classify videos and actions by producing a weighted mixture of augmented trajectories. Another major contribution by Wang et al. [96] was to use dense trajectories which were sampled by taking dense points from each frame. The dense optical flow field is used to calculate displacement for tracking dense trajectories with the calculation of other local descriptors (i.e., HOG, HOF).

Space–time local features Object recognition inspired the concept of using local features for action recognition from images, whereas local features are based on interest points and provide distinct features, which can be learned as features. The local features can be sparse (Harris 3D [12] and Dollar detector [97]) or dense (i.e., optical flow) depending on their extraction purpose. Jones et al. [98] have extended Dollar detector [97] using k-means for clustering of detected interest points and asymmetric bagging with random subspace support vector machine to incorporate feedback process. Gilbert et al. [99] have extended Harris 3D detector [12] to handle the sparsity issue and used hierarchical grouping for action classification. Sadek et al. [100] have used the concept of taking temporal self-similarities using the fuzzy log-polar histogram on Harris 3D detector [12] to describe the local interest points which are further classified through SVM. Ikizler-Cinbis and Sclaroff [101] have worked on feature extraction of various objects and humans by incorporating optical flow and foreground flow. The extracted features are fed to multiple instances of learning frameworks (MIL) to find the locality of interest points. Minhas et al. [102] have used 3D dual-tree discrete wavelet transform (DT-DWT) for spatiotemporal feature extraction and affine SIFT for local feature extraction. They have used a hybrid combination of both to feed the feature values to an extreme learning machine (ELM).

Sequential approaches The sequential approaches are of two types i.e., exemplar-based, and state-based which are briefly described as follows:

Exemplar-based The exemplar-based approaches use a representation of human actions as a template containing a set of sequences of an action that can be compared with new incoming video sequences, so the contributions were made to compare such templates for the process of action recognition. Darrell et al. [103] proposed a Dynamic Time Wrapping (DTW) algorithm to recognize and handle the variations in an action. It is extended by Gavrila et al. [104] to perform gesture analysis through DTW algorithm along with 3D joint angle model. Veeraraghavan et al. [105] proposed another modification to the DTW algorithm, in which they used a time function to monitor the overall activity process. It distinguishes between activities that appear similar but differ, for example, pulling, pushing, throwing, and so on. Another useful method is principal component analysis (PCA) and singular value decomposition (SVD). SVD is used for representation of video data to extract features as eigenvectors [106]. Then Efros et al. [107] attempted to use motion descriptors and incorporated optical flow as the baseline of the model. Optic flow is used to track human activity mainly in public places and set a threshold of 30 pixels for normal person's height. Lublinerman et al. [108] have proposed a similar system that was limited in performance due to noise and requires background enhancement. Jiang et al. [109] have worked on the use of geometric models to recognize actions through postures. Lin et al. [110] have worked on video representation and used k-mean clustering to generate prototype sequences. They have generated a unique prototype for each video using the prototype sequence estimation approach. For prototype matching, the fast DTW algorithm was used, which resulted in increased computational efficiency.

State model-based approaches The second category of sequential approaches is the state-based model which uses the hidden states for the representation of actions. Yamato et al. [111] have used Hidden Markov Model (HMM) for video representation and action recognition. HMM is already being used for speech recognition and text classification. A modification by Starner et al. [112] based on HMMs was proposed that targets the American Sign Language (ASL). In this approach, each sign is stored as HMMs to generate a corresponding sequence of features. An issue with this approach is that ASL can describe limited number of actions. Then Vogler et al. [113] have worked on reducing the number of combinations of ASL by using a Parallel Hidden Markov Models (PHMMs). Bobick et al. [114] have used the state model by representing gestures as 2D-trajectory which helps in finding the locality of interest points, i.e., location changes of hand. Another study has been conducted by Oliver et al. [115] to propose a Couple of Hidden Markov Model (CHMMs) which overcome the limitation of traditional HMM and makes it possible to analyze the interaction between over two people. Along with HMMs, Dynamic Bayesian networks (DBN) are also used for human body gesture estimation and a lot of improvement has been made to analyze person-person interactions [116]. In [117], Coupled hidden semi-Markov model is used to track the duration of actions occurring within an event (sub-events). It models the representation of a person-person interaction but results in a lot of model complexities which compromised the performance of the model. Gupta et al. [118] have used the probabilistic model which helps in the extraction of context-based information to perform the analysis of actions and demonstrated better performance in the object recognition process. Moore et al. [119] have introduced the use of both HMM and Bayesian relations for object classification and motion detection with limitation of hand moment detection only. Yu et al. [120] have presented another study that is based on the modification of HMMs. It has used star skeletons for representation to analyze the edges and corners of human postures through the application of contour and histogram-based methods. The novel texture descriptors were also being proposed by Kellokumpu et al. [121] for motion analysis with the use of HMM to assess the temporal information of motion histograms. Another work by Shi et al. [122] has been presented to resolve the inference issue while performing segmentation and recognition of human actions and proposed a dynamic programming algorithm (Viterbi like an algorithm) to perform action recognition.

Appearance-based approaches The appearance or outlook of any target can be presented through 2D (XY) and 3D (XYZ) depth images and such methods rely on the information related to shape, motion, and blend of both. Such methods use appearance-based information along with any suitable feature extraction method that can be shape and contour-based features and optic flow in case of motion-based features.

Shape-based approaches The human silhouette [123] is used to extract the local features, which are done by using foreground silhouette subtraction using a segmentation technique. The image can be assumed to have two spaces, i.e., positive space (image silhouette) and negative space (surrounding region between boundary of image and human) [124]. To work with human silhouette, one must use contour points, geometric information, and region-based features of frame, and a successful contribution was made to perform region-based feature extraction through division of human silhouette into fixed number of cells and grid to represent actions. The method further used a combination of two popular classifiers, support vector machine and Nearest Neighbor (SVM + NN) to recognize actions [125]. Another research was focused on considering the time-series data to use Symbolic Aggregate approximation (SAX) which first converts the silhouette into time-series data to produce SAX vector through applying random forest algorithm for action recognition [126]. Along with silhouette, pose invariant data are useful to estimate the actions through shape of human body, and a contour-based method is used, employing multi-view key poses for action recognition [127]. It is further extended through extraction of contour points from silhouette with radial scheme to perform action representation and classification through SVM [128]. Another method based on pose related information was proposed that uses scale invariant features from silhouettes. Key poses are produced through clustering of these features, which are fed to the weighted voting scheme for action recognition [129].

Motion-based approaches The basic trend is to extract the motion features through any useful mechanism and then apply a classifier to recognize actions. Such a contribution was made in [130] to produce a motion descriptor that uses motion directions and motion-intensity histograms of a moving body. Classification of different action categories is performed using SVM. Besides the motion descriptors, motion history images and histogram of oriented gradients (HoG) are also useful measures. Another useful approach was proposed in [131] to use the templates of motion which are based on motion history image and HoG. In [132], optic flow feature descriptor is used for human activity recognition and only motion-based features are extracted.

Hybrid approaches The hybrid approaches are based on the combination of both shape-based and motion-based information, such as an optic flow with silhouette-based features to perform view-invariant action recognition. In [133], also incorporated dimensionality reduction using principal component analysis. Another method was proposed to perform view invariant action recognition by using coarse silhouette with radial grid-based features and employing motion features [134]. Among these methods, in a study [135], action representation was done in the form of a sequence of the prototype by combining both motion and shape space. The action recognition of such representations is done by applying distance measures for sequence matching. The idea of combining both shape- and motion-based information was more improved by using motion energy images (MEI) and motion history image (MHI) for action key poses and action recognition is performed through nearest-neighbor classifier [136].

4.3.1.2 Hierarchical approaches The second feature-based category is hierarchical approaches, which have a lot of similarities with non-hierarchical approaches, especially for atomic actions. The hierarchical approaches mainly use complex activities by considering the sub-events within it, i.e., fighting, which involves other subtasks like pulling, pushing, punching, etc. Such approaches show their significance where flexibility is required while dealing with complex interactions, e.g., human–human interaction, human–object interaction. The hierarchical approaches are of three types which are statistical, syntactic, and description-based approaches.

Statistical approaches Initially, most of the statistical approaches were based on the extension or modification of Hidden Markov Model (HMM) and Dynamic Bayesian Networks (DBN) to handle concurrent and sequential sub-activities, respectively. After that, another hierarchical approach was proposed to emphasize the use of propagation networks (p-net), and these networks were proved significantly better for both sequential and concurrent activities [137]. Along with p-nets, a 4-layered probabilistic latent model [138] was proposed, which uses the Bayesian model for clustering after spatiotemporal feature extraction, and then recognition is performed through probabilistic latent model. The proposed model aimed to handle the atomic actions through clustered space–time features and complex actions with hierarchical descriptions. In another research, hierarchical clustering was proposed for action recognition through the representation of feature cues [139]. The cascade Condition Random Fields (CRFs) are helpful while analyzing the motion pattern, and SVM can classify these motion patterns as human actions [140]. Another research was conducted when data-related issues were raised and integration of training data with domain knowledge was proposed to resolve the insufficient data problem [141].

Syntactic approaches The activity is made up of multiple sub-activities and atomic actions which can be recognized by any activity recognition approaches such as Context-Free Grammar (CFGs)-based methods which are categorized under syntactic approaches. If the atomic sub-activities are symbols, then syntactic approaches integrate these in the form of a string of symbols but involve concurrent action recognition problems. To overcome the concurrent action recognition problem, a lot of improvements using CFGs are made, for instance, Stochastic CFG (SCFGs) used in [142, 143]. Activity recognition is performed by processing basic actions at lower layers of the model, while complex activities are recognized by applying parser techniques at top layer of the model. In [144], a method is proposed to handle the production rules problem, which means rules should be defined earlier. The proposed algorithm has done the task through automatic learning of rules Along with 2-layered frameworks, few researchers have put their efforts into producing multi-layer frameworks such as a 4-layered framework which uses the spatiotemporal features to generate a relevant set of rules for actions, i.e., strong, weak, and stochastic [145].

Description-based approaches The method, which can explicitly retain spatiotemporal structures extracted from human activities, is known as a description-based approach. Due to their explicit ability to describe the structure of spatiotemporal changes, description-based approaches can recognize both concurrent and sequential human activities. These methods use spatiotemporal and logical relationships to define relationships between simple actions that result in higher activity, such as sub-events. The CFGs with the use of formal syntax have been proposed for activity recognition [146, 147] and a PNF network for distinct temporal identification is used [148]. The famous Bayesian Belief Networks (BBN), event logic, and Petri nets have also been introduced for the task of complex activity recognition [149–151]. The Markov Logic Networks (MLN) that are symbolic were also proposed to conjecture human activities based on different probabilities [152]. Afterward, another postulation was proposed to handle higher level activity recognition by using different input sources based on temporal information employing no kind of probabilistic computation [89]. Another study [153] has attempted to perform event annotation of one-to-one basketball videos through mixed probabilistic and logical inference. The semantic description of different scenarios has been employed using first order logic to extract spatiotemporal knowledge and for basic information extraction, MLN is used.

The popularity of feature-based methods is increased through continuous improvement in existing methods, which includes techniques based on optical flow, Motion Boundary Histogram (MBH), Histogram of Oriented Gradients (HOG), Histogram of Optical Flow (HOF), and dense trajectories. Among these methods, IDT [154] remained a successful method, which is further changed by using Fisher vector for effective action recognition [155]. The space–time volume representation does not support view-invariant scenarios and is only useful when multiple people are involved in a single event. The space–time trajectories perform better with known video points and can accommodate different viewing angles. The space–time features can recognize multiple activities, but not complex activities with view-invariant representations. The appearance-based approaches primarily focus on using shape and motion-related information to generate motion descriptors or key poses for sequence matching. These methods use silhouette and interest point detectors, which are then fed into any suitable classifier (e.g., SVM) to perform action classification. Sequential approaches can deal with view-invariant data and complex activities. When compared to state-based methods, exemplar-based methods are more adaptable to complex activities and require less training data. Among layered approaches (i.e., hierarchical), description-based techniques outperformed other methods in terms of high-level activity recognition because of their explicit nature of maintaining spatiotemporal changes. Syntactic and statistical approaches have proven useful in dealing with noisy data.

4.3.2 Learning-based approaches

Learning-based approaches are the second major category of method used for HAR and Fig. 12 shows the general framework used by learning-based models to perform the task. Learning-based methods rely on automatic feature generation which does not require manual feature engineering process. Methods in this category have proven to be effective for a variety of tasks and can be used independently or in any hybrid combination (e.g., with handcrafted feature-based method). The learning-based approaches are subdivided into two main types i.e., non-neural networks-based approaches and neural networks-based approaches.

4.3.2.1 Non-neural network-based approaches *Genetic programming* The non-neural network-based approaches use the pre-defined set of rules or sequences for learning a model to evaluate the future data. The genetic programming and dictionary-based approaches are examples of such methods which are explained as follows: The genetic programming (GP) [156] is based on the Darwinian theory of selection and is famous for the vision-based tasks involving natural and random selection of solution set. The GP algorithm is an evolutionary algorithm that uses biologically inspired operators, i.e., crossover or mutation to perform the natural selection process over initialized computational program which should be randomly assembled. Shao et al. [157] have presented a study for evolution of motion features based on colors and optic flow fields by using the population of operators, i.e., 3D-Gabor filter [158] and wavelet [159]. Then classification error is calculated using GP fitness function along with SVM. The error is incorporated in the evolutionary algorithm that provides the final solution set in the form of cascaded operators for feature extraction process.

Dictionary learning Dictionary learning tries to learn the sparse data from the input by using linear combinations based on the dictionary atoms and these representations are useful mainly for categorization of data such as classification of images, detection, and action recognition. The bag-of-words model (BoW) [160] is very popular among researchers because of its usability and it is also based on dictionary learning. Guha et al. [161] have presented a method that overrules the performance of BoW model through sparse coding and improved the action recognition performance. The primary function of sparse representation is based on two components, i.e., regularization term and reconstruction-error value. Another concept of cross-view dictionary is proposed by Zheng et al. [162] to deal with the sparse coding and cross-domain inconsistencies so that action recognition can be view-invariant. Along with sparse coding, supervised dictionary learning-based approaches are also popular. In [163], a loosely supervised dictionary learning technique has been proposed to help the learning adaptation process from one action recognition dataset to another. It used both discriminative and cross-domain discrepancy terms to ensure smoothness in action recognition.

4.3.2.2 Neural network-based approaches Such types of methods try to model the human visual system to analyze data. A similar structure is used to build a learning model as biological neurons. Deep learning models have proven their worth in almost every domain where high-level data abstraction needs to be modeled and a few of them are discussed below:

Multi-layer perceptron network (fully connected neural networks) This type of method requires building the fully connected neural networks (FCNN) framework by using low-level information. Kim et al. [164] have designed a framework for feature extraction and classification by using FCNN as baseline and named it as Modified Neural Network (MNN). At first, handcrafted features

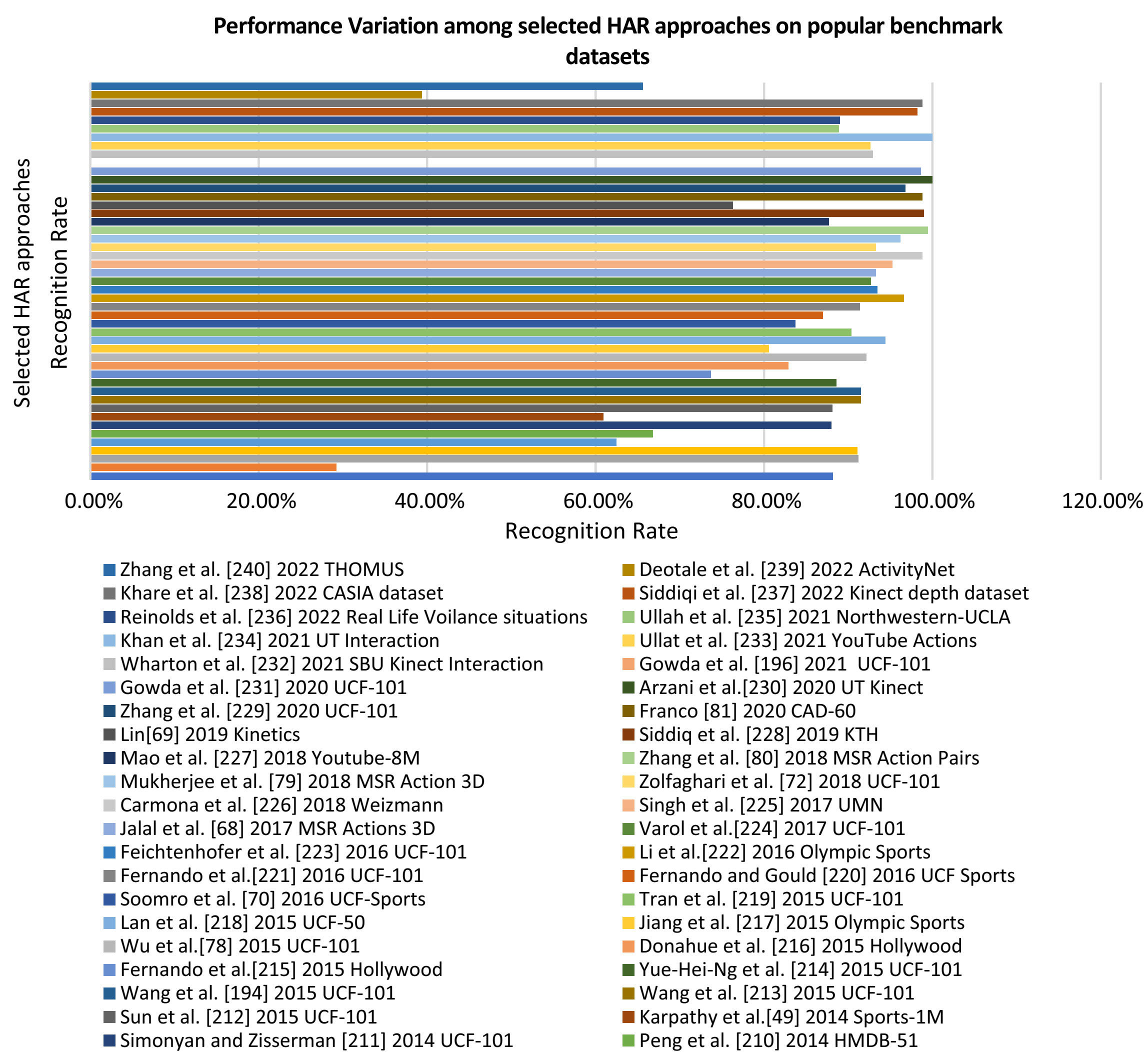

Fig. 13 Recognition rate variation in selected HAR approaches on different benchmark datasets

are used to extract the basic information, and then 2D contour of actor is obtained to generate spatiotemporal volume to obtain outer boundary information through 3D Gabor filters [165]. In [166], neural network applies to extract the action-based features which are done using four layers (i.e., two layers of convolutional and two sub-sampling layers) and a discriminative classifier to classify actions. Jhuang et al. [167] used layers based on spatiotemporal feature detectors through the approximation of motion directive units. It performed the global max computation of every feature map extracted from action data. Shao et al. [168] have proposed to use the multi-layer network instead of Restricted Boltzmann Machines (RBM) [169] to provide a hierarchical parametric network (HPN) using skeleton features. It has outperformed [170] to perform emission probability estimation of HMM.

Convolutional neural network (CNN) The 3D-CNN [171] is developed to perform the convolution on both spatial and temporal dimensions to extract features from multiple channels to provide a variety of action representations, whereas 2D-CNN [172] is only concerned with convolution on spatial domain. The proposed 3D-CNN model is a 7-layered architecture, including input layer, 3-convolutional, 2 sub-sampling, and a fully connected layer. Its feed-forward nature made it possible to extract features for action recognition. Another group of researchers [173] has presented Hierarchical Invariant spatiotemporal (HIST) framework. They have used Independent Subspace Analysis (ISA) [174] for feature extraction and Principal Component Analysis (PCA) [175] during the training to cater to large video data. So, the proposed HIST model works on training of multiple ISA, which is subsequently reduced through PCA and therefore uses the characteristics of ISA to perform unsupervised analysis, which can help label large video data. Then Baccouche et al. [176] have presented sequential deep learning model by using 3D-CNN and it works differently than the model presented in [164] because of its sequence of layers, i.e., two alternative convolutional-layers, rectification layer, sub-sampling layer, and another convolutional-layer, sub-sampling layer, third convolutional-layer, and then two fully connected layers. The action representations are extracted through CNN by capturing the temporal information over time with adaptation over sequential information and hence a sequential approach is used for action labeling.

Another variation is to use static frames as input for the action recognition process, such as OverFeat [177] or Caffe [178] have used image recognition model to learn static frames. Along with image-based frameworks, there are many video recognition frameworks that can recognize human activities from videos. Ning et al. [179] have proposed a video framework that can decompose videos into 2D images and then used the 2D CNN to analyze different stages of embryonic development. Later on, Karpathy et al. [51] have performed a comparative analysis to provide the best fit CNN-based architecture by using motion relative information on static videos. The experimentation involves the performance evaluation of different approaches on a famous large-scale action recognition dataset (i.e., YouTube-1 M) which reveals that the fixed-size architecture is not a suitable option for action recognition. Karpathy et al. [51] have presented 3D-CNN-based learning strategies, i.e., early fusion, late fusion, and slow fusion. Early fusion approach works by modifying 2D Convolutional window by adding temporal dimensions and passing these 3D cubes to first convolutional layer. In [180], CNN-Bi-LSTM is trained on RGB images to extract temporal information from video data. Late fusion strategy is applied at decision level of the network to provide end-to-end learning. The late fusion is based on incorporating two CNN on two distant frames then combines both at fully connected layer to extract the motion-related information at global level, whereas slow fusion is based on connecting two frames in both spatial and temporal dimensions.

Recurrent neural networks (RNN) Recurrent neural networks are popular for dealing with sequences because they act as a memory unit by storing the previous state and forwarding it to the next unit. They are computationally expensive, and RNN is further extended to reduce implementation issues, such as the Long Short-Term Memory Network (LSTM) and Gated Recurrent Unit (GRU). Yao et al. [181] have proposed the DT-3DResNet-LSTM to exploit the localities within video where any activity is taking place. The task is performed at different levels, such as first detected object becomes the input of object tracking model and that clipped video frame is fed to CNN for feature extraction. Then LSTM is used for HAR classification and final temporal information is achieved. Meng et al. [182] have proposed the Quaternion Spatiotemporal convolutional neural network (QST-CNN) and Long Short Term Memory network (LSTM) which is known as QST-CNN-LSTM to use on RGB data by considering its spatiotemporal information. LSTM is used to capture the difference between two frames of a video. The model works through motion region extraction and this outperforms both on UCF11 and UCF sports datasets. Another work [183] performs group activity recognition, and the proposed model is known as stagNet which uses Spatiotemporal and semantics of the data to feed RNN. This makes the model learn the intergroup relation and

Table 4 Limitations within HAR

Limitations	Description
Anthropometric variation [47, 55, 69, 127, 133, 136]	Anthropometric issues are related to postures and angle issues that arise primarily because of human body variation and acting in various poses. Such problems can occur when using shape or pose feature extraction
Multiview variation [55, 101, 128, 130, 131, 133]	Multi-view variation occurs because of unsettled camera view, which gives a different perspective to any actions. To avoid confusion, use synchronized multi-camera while producing features from each view
Cluttered and dynamic scene variation [15, 101, 211, 243]	It applies to recorded action datasets in which actions are recorded in an indoor environment to provide static and uniform background throughout the activity, but such an approach causes problems when we evaluate these methods in outdoor conditions with highly dynamic backgrounds. Many activity recognition algorithms, such as the optic flow method, combine background noise with human motion information to overcome these issues
Intra-class variability and inter-class similarity [79]	Intra-class variability and inter-class similarity arise from the unique behavior of each human being and the tendency to repeat the same actions. For example, everyone walks differently due to age or muscular condition, so we must deal with complex scenarios. We cannot rely on a single model to perform the same task, so such issues must be addressed, i.e., by using discriminative features
Occlusion [5, 10, 55, 101]	The occlusion problem occurs when the human body is obscured by another frontal object, which can be caused by self-interference of different body parts or by another object
Environmental constraints [55, 243–245]	The light effect changes the overall impression of a scene. The light sources cast shadows on the objects and cause variations in illumination. Changes in weather and daytime conditions cause a significant change in the scene and the created artifacts; for example, an action recorded during rain differs completely from the same action recorded in broad daylight or the evening
Dynamic Cameras [55, 101, 246]	The HAR is relatively simple in static camera scenarios, but the variations introduced by dynamic cameras cause changes in pose and illumination, making the problem more difficult
Inadequate data [53, 199, 247]	While working with deep learning models, the amount of data are the most important consideration. The limitation in data amount is primarily because of the difficulty in creating and labeling human activity videos, as well as processing and storage limitations in some systems

distinguishable spatiotemporal features of the data. The stagNet is further extended to provide group activity and individual action recognition by incorporating body regions and global part feature pooling [184]. The authors in [185] have described pre-trained weights may affect the learning of a model and it can be addressed through a bi directional long short-term memory (BiLSTM) model. They have proposed to use an attention mechanism to prioritize the human actions from a video sequence. The authors in [186] have used dilated CNN (DCNN) layers for feature extraction and BiLSTM used these features to analyze long-term dependencies. The process of action recognition has been further improved by applying attention mechanism, which can extract high-level patterns. The authors have proposed densely connected Bi-directional LSTM (DB-LSTM) to improve the robustness of the model. Their model works in both forward and backward direction to model visual and temporal information of data. Moreover, authors have used appearance and motion-based modalities to improve the human activity recognition. In [187], authors have proposed spatial–temporal differential long short term memory (ST-D LSTM) and used Inception V3 for feature extraction from video data. The features fed to the enhanced input differential module and spatial memory state module. Spatial information is extracted and transferred horizontally and the data are forwarded to traditional long-term convolutional networks to evaluate the performance of proposed model.

Transformer networks Transformer networks [188] are popular among both natural language processing and computer vision tasks. They are used to model sequential data input, such as audios and videos, like recurrent neural networks. Transformer network is mainly used for speech recognition, but it is also popular among action recognition tasks such as multimodal-based methods. Deep learning-based methods are usually complex and require many parameters to train model, which increases the computational overhead. Transformer networks can produce a less amount of trainable parameters. Video transformer network [189], is a sequence-based neural network architecture that attempts to recognize long range dependencies and analyze full length videos. The authors have claimed that their model works better than any 2D spatial network and achieved faster training time with fewer GFLOPs (Giga floating point operations). In [190], authors have used Spatiotemporal-based transformer networks model (ST-TR), which uses Skeleton-based data. The authors have used sparse attention mechanism on spatial information of human actions to extract intra-frames interactions of different body parts. They have used temporal self-attention mechanism to produce inter frame correlations which help to model skeleton data for action recognition. Action Transformer [191], is fully self-attentional architecture which has performed better than complex CNN, RNN and attentive layer-based architectures. The authors have worked on pose representation through small temporal window which have produced a low latency overhead for accurate recognition. They have also published a large-scale dataset entitled "MPOSE2021" which can be used for real-time, short-time human action recognition.

Auto-encoder Auto-encoders learn data representation through unsupervised learning, primarily for low dimensionality. The authors of [192] used auto-encoders with CNN to perform online action recognition, with CNN learning frame-level representation. As a result, the auto-encoder performs sequence learning and feature dimensionality reduction. Then, unlike CNN, another group of researchers used such methods for HAR and auto-encoders for abnormal activity detection, which learns spatiotemporal features from data to avoid missing label problems [193].

Generative adversarial network (GAN) Generative methods use an unsupervised learning regime to learn data representation from any type of unlabeled data. These methods are popular for generating synthetic data, which is achieved by learning features of each class from original data. Today, we have a large amount of unlabeled data that are useless without labeling, but generative methods have made it possible to work with such data. A group of researchers has used GAN for early prediction of human activity in which GAN is used to avoid motion blur problems and predict future motion [194].

Hybrid model Hybrid models are based on using handcrafted features, along with neural network models, to use the benefit of both strategies. Simonyan et al. [195] have proposed a 2-stream CNN-based architecture through the decomposition of video data into both spatial and temporal domain and then a CNN is trained on top of optic flow. The authors have proposed a lot of variations such as optic flow stacking, trajectory stacking, and bi-directional optic flow while 2-stream training is performed on HMDB-51 [54] and UCF-101 [45] datasets to compare the classification accuracy. The proposed architecture is a hybrid model because of the use of a CNN model and learning from both handcrafted features and raw pixels. The 2-stream CNN architecture is extended by Wang et al. [196] by introducing the use of trajectories along with it. Then 2-stream trajectory pooled deep convolutional descriptor (TDD) [154] has been proposed, which has also been trained on HMDB-51 and UCF-101 datasets to provide a generalized feature extractor for future videos. In [197], authors proposed the use of dense trajectories and discriminative Fisher vector to encode TDDs via fisher vector representation.

4.3.2.3 Learning strategy *Transfer learning* Transfer learning is a type of learning in which learning from one network is transferred to another network in terms of weights to improve recognition results. There are several transfer learning strategies, such as freezing the convolutional layer of a new network and allowing only fully connected layers to perform classification of tasks where the target problem is like a pre-trained model (To perform Sports activity recognition, pre-trained model of Sports-1 M can be used). Du et al. [199] proposed a cascaded architecture for activity recognition that is based on a convolutional neural network.

Zero-shot learning Zero shot learning is a type of learning where we deal with unseen classes of data, normally with synthetic data generation. Gowda et al. [198] have proposed a reinforcement learning model to learn all classes at once rather than individual data points optimization. Moreover, [199, 200] have used zero action recognition and, [82] proposed a semantic-based multi stream deep neural network that learns both action and action attributes.

In Short, the models presented in [195] and [196] have been built on using handcrafted features-based methods using convolutional architecture as a baseline. The framework proposed in [201] also uses the convolutional architecture to learn the motion related actions. And learning-based approaches can be discussed as to how and when learning framework is used because only few studies are based on direct use of CNN and the rest follow a hybrid regime. The learning process works entirely differently if the two frames are fused by following different learning strategies, i.e., early, late, and slow fusion. The learning frameworks may vary because of the number of layers within a network, such as slow fusion CNN has maximum number of layers. Some note that performance of slow fusion CNN (SFCNN) is not satisfactory while comparing the feature-based shallow representations [154, 155]. This means greater number of layers do not promise better results. Two other deep learning frameworks, early fusion CNN [195] and late fusion CNN [196] have initially performed well, but both have resulted in reduced performance while being tested on spatial stream networks. While working with CNN-based networks, the major problem is the size of dataset as majority of datasets have small number of representative videos with missing labels. For training, two datasets can be combined to increase the data volume, but because of intersection of different action classes, it is not a suitable option. Along with discussed methods, multi-task [202–204] and transfer learning [205–207]-based approaches are also in use, which helps in combining the data or to use the learned representation of one dataset with another dataset. For example, Wang et al. [196] has used the transfer learning through training the model on UCF-101 and then trained on HMDB-51 to extract features for action classification.

4.4 Analysis of State-of-the-art HAR approaches

We have discussed HAR approaches in the above section, followed by taxonomy to cover online/offline processing based, modality based, and method-based approaches. In this section, state-of-the-art methods are analyzed to provide recent trends and to highlight domain challenges and we have performed our analysis on 46 state-of-the-art techniques of HAR. The selected techniques include machine learning approaches, deep learning approaches, multimodal approaches, and a framework for online HAR. We have analyzed all selected studies using six major parameters, which are publication year, method type, data input, activity level, dataset size, and its performance on benchmark datasets. In this study, publication year is important for signifying a study because recent methods are close to general trend of HAR. The second parameter is type of methods used to perform recognition, which demonstrates the popularity of specific type of method among both handcrafted and learning-based HAR. Then data input represents if the videos are stacked in a database or based on real-time camera feed. Activity level is another parameter that shows a method is useful for recognition of simple, intermediate, or complex activities. All studies include experiments on relevant HAR benchmark datasets which may help new researchers to identify which datasets are more useful depending on problem. Moreover, size of dataset is relevant to the type of method used, for example, deep learning-based solutions perform well with large datasets and handcrafted feature-based methods with small or intermediated sized datasets. Performance of methods is important parameter too and we have mentioned achieved performance of all selected studies. Table 3 provides quantitative analysis of HAR approaches which is based on above-mentioned parameters.

4.4.1 Discussion

Table 3 is based on 46 state-of-the-art methods from 2011 to 2022, 20 of which are handcrafted and the remaining 26 are learning-based approaches. This table provides details of recent methods; popular datasets used for different tasks, and achieved performance. Performance does not directly affect importance of a study as not all studies are focused on increasing recognition rate. Studies published earlier were more focused on increasing recognition rate, but later on a lot of challenges are highlighted and researchers start working on multiple perspectives of a domain. For example, Convolutional neural networks-based methods. Figure 13 provides a performance graph that shows how the activity recognition rate changed over a decade based on HAR benchmarks. We have added performance of studies in “average precision” column, which provides performance on different benchmark datasets. For example, Table 3 shows that 2011 has performance on two datasets only, whereas 2020 shows performance on eleven datasets. Therefore, the graph can be dense at places depending on the number of datasets used in the following year by selected set of studies. UCF-101 and HMDB-51 both are highly cited datasets and Table 3 also presents that most of the researchers have performed their experiment on these two.

Table 3 shows that in [233] authors have achieved a good performance on UCF-101, HMDB-51, and ActivityNet by using temporal segment network-based approach. Later in 2021 [198], authors have tried to improve the unseen class recognition problem by using zero-shot action recognition. They achieved a low performance value as compared to [233] as it was focused on unseen class recognition, which means they have used UCF-101 and HMD-51 as training data only. For performance evaluation, data are randomly taken from any action class, which resulted in low performance when we compare it to other mentioned studies. Unseen class-based action recognition is still an open research area and needs a lot of improvement. HAR is a diverse domain that includes simple to complex activities and among these intermediate activities are frequently recognized in the selected studies. The task complexity depends on type of activity recognized by a study, for example, daily actions are simple tasks, whereas group activities or crowd behavior is a complex task. Human-to-human and human-to-object interactions are categorized as intermediate tasks. Another variation is the size of the dataset, which has a significant impact on the recognition rate and is directly related to the type and level of activities. Among HAR, abnormal activities still need improvement as they may be affected by various factors such as Reinolds et al. [238] have proposed that abnormal activities are influenced by audio too. They have performed recognition by extracting both audio-based features and video-based features to perform recognition. They also compared features of both and claimed that video-based features are more useful to perform recognition. In Table 3 since 2020, nine methods are based on deep learning-based approaches, whereas only five are from handcrafted features-based approaches, which shows an inclined behavior of researchers toward deep learning. This is because, due to technological advancements, large data are available in form of videos. However, it still needs a good annotation method, and hybrid approaches (combination of both handcrafted features-based and deep learning-based approaches) are getting attention due to their performance. Table 3 shows that most of the studies are based on using intermediate size of dataset for experiment as small size datasets have a limited number of training instances. Small size datasets may compromise model performance, whereas large size datasets are computationally expensive to deal with. However, large datasets provide better learning and hence provides better performance. Therefore, if the resources are not a bottleneck, it is better to use large size datasets for human activity recognition. In [208], for example, a handcrafted feature-based approach that is trained on a small dataset is used to perform simple action recognition. Datasets must be chosen based on the task and method, for example, large-scale datasets are more popular among deep learning solutions. It should not be dependent, but current resources and research need to be improved to deal with both data size and data type issues. GAN is widely used, and most researchers are working on synthetic data generation to cover potential HAR scenarios.

5 Limitations and future research

The preceding discussion can be expanded to highlight the limitations of datasets across HAR variables. Table 4 includes the pertinent details to present the highlighted issues while considering state-of-the-art approaches.

5.1 Future research directions

HAR is constantly evolving and offers promising performance ranging from simple day-to-day living systems to real-time surveillance systems. Its multi-purpose application has made it an ever-active research area, and with each technical advancement, new research directions are opened. Hence, It is essential to design a representative dataset for HAR that can overcome the occlusion, view variation, and weather constraints of recorded data.

5.1.1 Improvement within benchmark datasets

The preparation of data and approaches for multi-camera-based human activity recognition also needs to be improved. The size of the dataset and activity classes with proper labels is another important consideration. For example, YouTube 8-M is the largest dataset with a variety of classes, but not all activities are covered. Collecting a large-scale dataset of human activities, either by combining existing datasets or by adding new samples, could solve this problem. However, this may necessitate time-consuming labeling of the content and its temporal position. To avoid the time-consuming manual process, another option is synthetic data generation and synthetic label generation. As some activities occur with relatively few anomalies, class imbalance is also an open issue. Normally, data augmentation is used to solve the problem of class imbalance, but synthetic data generation is also an option. Because human activity data contain subtle variations, synthetic data generation necessitates further investigation. Another approach is to use a weakly supervised learning strategy with web-based videos that have weak labels. This may solve the problem of small dataset size and improve the overall performance of HAR system.

5.1.2 Improvement in models

Deep learning has proven its worth everywhere, including HAR. However, deep learning models are improving all the time, and another improvement can be made by modifying the global average pooling layer in existing 3D deep convolution neural networks. Using temporal information or Improved Dense Trajectories (IDT) may be useful for this purpose. Multimodal approaches rely on the fusion of data from various devices, such as audio-visual data. Such information can be more useful in distinguishing visually similar activities. Researchers can extend HAR to improve performance of traditional ML approaches on large-scale datasets. Normally, HAR is performed on large videos, which may have irrelevant frames and are not part of the recognition. Machine learning models require improvement to identify trimmed actions from large videos. Deep learning models such as convolutional architecture-based models (3D CNN) can be extended to exploit spatiotemporal information of action data. We can improve HAR generalization problem by improving reinforcement and active learning strategies. Another problem that needs improvement is to classify overlapped actions. Therefore, classification models can also be improved for classifying overlapped actions from the dataset. Multimodal approaches require improvement to perform fusion of multiple modalities to perform HAR. Data from multimodal sources can also be exploited to perform action recognition along with the emotional state of human (e.g., walking in anger, running in fear, smiling while talking). Multimodal approaches improved for human activity recognition with emotional state identification may help to improve context-based human activity recognition process. Similarly, another approach to augmenting visual data for better learning is to use multi-camera views and data fusion from heterogeneous devices. Cross-domain transfer learning and deep learning models of multimodal data can be an interesting dimension to explore [83].

6 Conclusion

This survey presents various aspects of video-based human activity recognition to provide an up-to-date and generalized perspective as compared to previous surveys. It considers model-based, modality-based, and online/offline setting-based variation in HAR approaches. Our survey explains type of activities, task complexities, benchmark datasets, and analysis of state-of-the-art approaches to highlight HAR trends. Over 30 HAR datasets, 46 state-of-the-art approaches, and 20 state-of-the-art surveys are discussed in survey. This study presents a taxonomy that incorporates a variety of methods useful for online/offline, unimodal/multimodal, and handcrafted feature-based approaches/learning-based approaches. This study also includes benchmark datasets, as HAR performance equally depends on choice of datasets. We have categorized the datasets into single-view, multi-view, RGB, RGB-D, activity types, and application areas. This study shows intermediate size datasets are popular among researchers because of the trade-off between accuracy and resources. This survey provides a comparative analysis of recent methods that shows both handcrafted and learning-based approaches are improving consistently. However, learning-based models and hybrid models became popular because of availability of large amount of data. Multimodal approaches and online HAR approaches have gained popularity because of their strengths, but both have room for improvement. Multimodal approaches use multiple data cues, which make it computationally expensive, whereas online methods are expensive because of the online processing of video frames. Hence, our survey covers multiple dimensions of HAR to give a complete overview, including methods, datasets, challenges, and future directions.

Acknowledgements We acknowledge partial support from the National Center of Big Data and Cloud Computing (NCBC) and Higher Education Commission (HEC) of Pakistan for conducting this research.

Data availability Data sharing is not applicable to this article as no datasets were produced or analyzed during the current study. However, this study is based on analyzing existing methods, and their sources are added to the manuscript.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

References

1. Beddiar DR, Nini B, Sabokrou M, Hadid A (2020) Vision-based human activity recognition: a survey. Multimed Tools Appl 79(41):30509–30555
2. Huang S-C (2010) An advanced motion detection algorithm with video quality analysis for video surveillance systems. IEEE Trans Circuits Syst Video Technol 21(1):1–14
3. Cheng F-C, Huang S-C, Ruan S-J (2010) "Scene analysis for object detection in advanced surveillance systems using Laplacian distribution model. IEEE Trans Syst Man Cybern Part C 41(5):589–598
4. Oral M, Deniz U (2007) Centre of mass model–a novel approach to background modelling for segmentation of moving objects. Image Vis Comput 25(8):1365–1376
5. Yilmaz A, Li X, Shah M (2004) Contour-based object tracking with occlusion handling in video acquired using mobile cameras. IEEE Trans Pattern Anal Mach Intell 26(11):1531–1536
6. Bernardin K, Stiefelhagen R (2008) Evaluating multiple object tracking performance: the CLEAR MOT metrics. EURASIP J Image Video Process 2008:1–10
7. Bolme DS, Beveridge JR, Draper BA, Lui YM (2010) Visual object tracking using adaptive correlation filters. In: 2010 IEEE computer society conference on computer vision and pattern recognition, IEEE, pp 2544–2550
8. Cucchiara R, Grana C, Piccardi M, Prati A (2003) Detecting moving objects, ghosts, and shadows in video streams. IEEE Trans Pattern Anal Mach Intell 25(10):1337–1342
9. Denman S, Fookes C, Sridharan S (2009) Improved simultaneous computation of motion detection and optical flow for object tracking. In: 2009 Digital image computing: techniques and applications, IEEE, pp 175–182
10. Ince S, Konrad J (2008) Occlusion-aware optical flow estimation. IEEE Trans Image Process 17(8):1443–1451
11. Morris BT, Trivedi MM (2008) A survey of vision-based trajectory learning and analysis for surveillance. IEEE Trans Circuits Syst Video Technol 18(8):1114–1127
12. Laptev I (2005) On space-time interest points. Int J Comput Vision 64(2–3):107–123
13. Blunsom P (2004) Maximum entropy markov models for semantic role labelling. Proc Australasian Lang Technol Workshop 2004:109–116
14. Nunez JC, Cabido R, Pantrigo JJ, Montemayor AS, Velez JF (2018) Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recogn 76:80–94
15. Chen X, Guo H, Wang G, Zhang L (2017) Motion feature augmented recurrent neural network for skeleton-based dynamic hand gesture recognition. In: 2017 IEEE international conference on image processing (ICIP), IEEE, pp 2881–2885
16. Li C, Hou Y, Wang P, Li W (2017) Joint distance maps based action recognition with convolutional neural networks. IEEE Signal Process Lett 24(5):624–628
17. Kerber F, Puhl M, Krüger A (2017) User-independent real-time hand gesture recognition based on surface electromyography. In: Proceedings of the 19th international conference on human-computer interaction with mobile devices and services, pp 1–7
18. Vishwakarma S, Agrawal A (2013) A survey on activity recognition and behavior understanding in video surveillance. Vis Comput 29(10):983–1009
19. Zhen X, Shao L, Maybank S, Chellappa R (2016) Handcrafted vs. learned representations for human action recognition. Image Vis Comput 55(2):39–41

20. Sargano AB, Angelov P, Habib Z (2017) A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. Appl Sci 7(1):110
21. Ke S-R, Thuc HLU, Lee Y-J, Hwang J-N, Yoo J-H, Choi K-H (2013) A review on video-based human activity recognition. Computers 2(2):88–131
22. Cheng G, Wan Y, Saudagar A, Namuduri K, Buckles B (2015) Advances in human action recognition: a survey. arXiv preprint arXiv:1501.05964
23. Dawn DD, Shaikh SH (2016) A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. Vis Comput 32(3):289–306
24. Vrigkas M, Nikou C, Kakadiaris IA (2015) A review of human activity recognition methods. Front Robot AI 2:28
25. Herath S, Harandi M, Porikli F (2017) Going deeper into action recognition: a survey. Image Vis Comput 60:4–21
26. Jegham I, Khalifa AB, Alouani I, Mahjoub MA (2020) Vision-based human action recognition: an overview and real world challenges. Forensic Sci Int Digit Invest 32:200901
27. Wang Z et al (2019) A survey on human behavior recognition using channel state information. IEEE Access 7:155986–156024
28. Rodríguez-Moreno I, Martínez-Otzeta JM, Sierra B, Rodriguez I, Jauregi E (2019) Video activity recognition: state-of-the-art. Sensors 19(14):3160
29. Liu J, Liu H, Chen Y, Wang Y, Wang C (2019) Wireless sensing for human activity: a survey. IEEE Commun Surv Tutor 22(3):1629–1645
30. Dang LM, Min K, Wang H, Piran MJ, Lee CH, Moon H (2020) Sensor-based and vision-based human activity recognition: a comprehensive survey. Pattern Recogn 108:107561
31. Chaurasia SK, Reddy S (2022) State-of-the-art survey on activity recognition and classification using smartphones and wearable sensors. Multimedia Tools Appl 81(1):1077–1108
32. Yao G, Lei T, Zhong J (2019) A review of convolutional-neural-network-based action recognition. Pattern Recogn Lett 118:14–22
33. Zhang H-B et al (2019) A comprehensive survey of vision-based human action recognition methods. Sensors 19(5):1005
34. Das B, Saha A (2021) A survey on current trends in human action recognition. In: Advances in medical physics and healthcare engineering, Springer, pp 443–453
35. Gupta N, Gupta SK, Pathak RK, Jain V, Rashidi P, Suri JS (2022) Human activity recognition in artificial intelligence framework: a narrative review. Artif Intell Rev 3:1–54
36. Zhu F, Shao L, Xie J, Fang Y (2016) From handcrafted to learned representations for human action recognition: a survey. Image Vis Comput 55:42–52
37. Tripathi RK, Jalal AS, Agrawal SC (2018) Suspicious human activity recognition: a review. Artif Intell Rev 50(2):283–339
38. Chaquet JM, Carmona EJ, Fernández-Caballero A (2013) A survey of video datasets for human action and activity recognition. Comput Vis Image Underst 117(6):633–659
39. Zhang J, Li W, Ogunbona PO, Wang P, Tang C (2016) RGB-D-based action recognition datasets: a survey. Pattern Recogn 60:86–105
40. Singh T, Vishwakarma DK (2019) Video benchmarks of human action datasets: a review. Artif Intell Rev 52(2):1107–1154
41. Wang J, Nie X, Xia Y, Wu Y, Zhu S-C (2014) Cross-view action modeling, learning and recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2649–2656
42. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004, vol. 3: IEEE, pp 32–36
43. Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. IEEE Trans Pattern Anal Mach Intell 29(12):2247–2253
44. Xia L, Chen C-C, Aggarwal JK (2012) View invariant human action recognition using histograms of 3d joints. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops, IEEE, pp 20–27
45. Soomro K, Zamir AR, Shah M (2012) A dataset of 101 human action classes from videos in the wild. Center Res Comput Vis 2:666
46. Rahmani A, Mahmood A, Huynh D, Mian A (2014) Action classification with locality-constrained linear coding. In: 2014 22nd international conference on pattern recognition, IEEE, pp 3511–3516
47. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vis Image Underst 104(2–3):249–257
48. Niebles JC, Chen C-W, Fei-Fei L (2010) Modeling temporal structure of decomposable motion segments for activity classification. European conference on computer vision. Springer, Berlin, pp 392–405
49. Marszalek M, Laptev I, Schmid C (2009) Actions in context. In: 2009 IEEE conference on computer vision and pattern recognition, IEEE, pp 2929–2936
50. Reddy KK, Shah M (2013) Recognizing 50 human action categories of web videos. Mach Vis Appl 24(5):971–981
51. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp 1725–1732
52. Heilbron FC, Escorcia V, Ghanem B, Niebles JC (2015) Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 961–970
53. Abu-El-Haija S et al. (2016) Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675
54. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: a large video database for human motion recognition. In: 2011 international conference on computer vision, IEEE, pp 2556–2563
55. Yu S, Tan D, Tan T (2006) A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th international conference on pattern recognition (ICPR'06), vol 4: IEEE, pp 441–444
56. Gu C et al. (2018) Ava: a video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6047–6056
57. Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6479–6488
58. Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3d points. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, IEEE, pp 9–14
59. Berclaz J, Fleuret F, Turetken E, Fua P (2011) Multiple object tracking using k-shortest paths optimization. IEEE Trans Pattern Anal Mach Intell 33(9):1806–1819
60. Hu J-F, Zheng W-S, Ma L, Wang G, Lai J (2016) Real-time RGB-D activity prediction by soft regression. European Conference on Computer Vision. Springer, Berlin, pp 280–296
61. Sung J, Ponce C, Selman B, Saxena A (2012) Unstructured human activity detection from rgbd images. In: 2012 IEEE international conference on robotics and automation, IEEE, pp 842–849

62. Koppula HS, Gupta R, Saxena A (2013) Learning human activities and object affordances from rgb-d videos. Int J Robot Res 32(8):951–970
63. Chen C, Jafari R, Kehtarnavaz N (2015) UTD-MHAD: a multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE international conference on image processing (ICIP), IEEE, pp 168–172
64. Ni B, Wang G, Moulin P (2011) Rgbd-hudaact: A color-depth video database for human daily activity recognition. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops), IEEE, pp 1147–1153
65. Ofli F, Chaudhry R, Kurillo G, Vidal R, Bajcsy R (2013) Berkeley mhad: a comprehensive multimodal human action database. In: 2013 IEEE workshop on applications of computer vision (WACV), IEEE, pp 53–60
66. Wolf C et al (2014) Evaluation of video activity localizations integrating quality and quantity measurements. Comput Vis Image Underst 127:14–30
67. Bloom V, Argyriou V, Makris D (2014) G3di: A gaming interaction dataset with a real time detection and evaluation framework. European conference on computer vision. Springer, Berlin, pp 698–712
68. Shahroudy A, Liu J, Ng T-T, Wang G (2016) Ntu rgb+ d: a large scale dataset for 3d human activity analysis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1010–1019
69. Van Gemeren C, Tan RT, Poppe R, Veltkamp RC (2014) Dyadic interaction detection from pose and flow. International Workshop on Human Behavior Understanding. Springer, Berlin, pp 101–115
70. Jalal A, Kim Y-H, Kim Y-J, Kamal S, Kim D (2017) Robust human activity recognition from depth video using spatiotemporal multi-fused features. Pattern Recogn 61:295–308
71. Lin J, Gan C, Han S (2019) Tsm: temporal shift module for efficient video understanding. In: Proceedings of the IEEE international conference on computer vision, pp 7083–7093
72. Soomro K, Idrees H, Shah M (2016) Predicting the where and what of actors and actions through online action localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2648–2657
73. Singh G, Saha S, Sapienza M, Torr PH, Cuzzolin F (2017) Online real-time multiple spatiotemporal action localisation and prediction. In: Proceedings of the IEEE international conference on computer vision, pp 3637–3646
74. Zolfaghari M, Singh K, Brox T (2018) Eco: efficient convolutional network for online video understanding. In: Proceedings of the European conference on computer vision (ECCV), pp 695–712
75. Xu M, Gao M, Chen Y-T, Davis LS, Crandall DJ (2019) Temporal recurrent networks for online action detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 5532–5541
76. Gao M, Zhou Y, Xu R, Socher R, Xiong C (2020) WOAD: weakly supervised online action detection in untrimmed videos. arXiv preprint arXiv:2006.03732
77. Ye Y, Li K, Qi G-J, Hua KA (2015) Temporal order-preserving dynamic quantization for human action recognition from multimodal sensor streams. In: Proceedings of the 5th ACM on international conference on multimedia retrieval, pp 99–106
78. Vrigkas M, Nikou C, Kakadiadis IA (2014) Classifying behavioral attributes using conditional random fields. Hellenic conference on artificial intelligence. Springer, Berlin, pp 95–104
79. Shahroudy A, Ng T-T, Yang Q, Wang G (2015) Multimodal multipart learning for action recognition in depth videos. IEEE Trans Pattern Anal Mach Intell 38(10):2123–2129
80. Wu Z, Jiang Y-G, Wang X, Ye H, Xue X, Wang J (2015) Fusing multi-stream deep networks for video classification. arXiv preprint arXiv:1509.06086
81. Mukherjee S, Anvitha L, Lahari TM (2018) Human activity recognition in RGB-D videos by dynamic images. arXiv preprint arXiv:1807.02947
82. Zhang C, Tian Y, Guo X, Liu J (2018) DAAL: deep activation-based attribute learning for action recognition in depth videos. Comput Vis Image Underst 167:37–49
83. Franco A, Magnani A, Maio D (2020) A multimodal approach for human activity recognition based on skeleton and RGB data. Pattern Recogn Lett 131:293–299
84. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267
85. Hu Y, Cao L, Lv F, Yan S, Gong Y, Huang TS (2009) Action detection in complex scenes with spatial and temporal ambiguities. In: 2009 IEEE 12th international conference on computer vision, IEEE, pp 128–135
86. Roh M-C, Shin H-K, Lee S-W (2010) View-independent human action recognition with volume motion template on single stereo camera. Pattern Recogn Lett 31(7):639–647
87. Qian H, Mao Y, Xiang W, Wang Z (2010) Recognition of human activities using SVM multi-class classifier. Pattern Recogn Lett 31(2):100–111
88. Kim W, Lee J, Kim M, Oh D, Kim C (2010) Human action recognition using ordinal measure of accumulated motion. EURASIP J Adv Signal Process 2010(1):1–11
89. Ijsselmuiden J, Stiefelhagen R (2010) Towards high-level human activity recognition through computer vision and temporal logic. Annual conference on artificial intelligence. Springer, Berlin, pp 426–435
90. Fang C-H, Chen J-C, Tseng C-C, Lien J-JJ (2009) Human action recognition using spatio-temporal classification. Asian conference on computer vision. Springer, Berlin, pp 98–109
91. Ziaeefard M, Ebrahimnezhad H (2010) Hierarchical human action recognition by normalized-polar histogram. In: 2010 20th international conference on pattern recognition, IEEE, pp 3720–3723
92. Wang Y, Mori G (2009) Human action recognition by semilatent topic models. IEEE Trans Pattern Anal Mach Intell 31(10):1762–1774
93. Guo K, Ishwar P, Konrad J (2009) Action recognition in video by covariance matching of silhouette tunnels. In: 2009 XXII Brazilian symposium on computer graphics and image processing, IEEE, pp 299–306
94. Kim T-K, Cipolla R (2008) Canonical correlation analysis of video volume tensors for action categorization and detection. IEEE Trans Pattern Anal Mach Intell 31(8):1415–1428
95. Messing R, Pal C, Kautz H (2009) Activity recognition using the velocity histories of tracked keypoints. In: 2009 IEEE 12th international conference on computer vision, IEEE, pp 104–111
96. Wang H, Kläser A, Schmid C, Liu C-L (2011) Action recognition by dense trajectories. In: CVPR 2011, IEEE, pp 3169–3176
97. Dollár P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: 2005 IEEE international workshop on visual surveillance and performance evaluation of tracking and surveillance, IEEE, pp 65–72
98. Jones S, Shao L, Zhang J, Liu Y (2012) Relevance feedback for real-world human action retrieval. Pattern Recogn Lett 33(4):446–452
99. Gilbert A, Illingworth J, Bowden R (2009) Fast realistic multi-action recognition using mined dense spatio-temporal features. In: 2009 IEEE 12th international conference on computer vision, IEEE, pp 925–931

100. Sadek S, Al-Hamadi A, Michaelis B, Sayed U (2011) An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity. EURASIP J Adv Signal Process 2011(1):540375
101. Ikizler-Cinbis N, Sclaroff S (2010) Object, scene and actions: Combining multiple features for human action recognition. European conference on computer vision. Springer, Berlin, pp 494–507
102. Minhas R, Baradarani A, Seifzadeh S, Wu QJ (2010) Human action recognition using extreme learning machine based on visual vocabularies. Neurocomputing 73(10–12):1906–1917
103. Darrell T, Pentland A (1993) Space-time gestures. In: Proceedings of IEEE conference on computer vision and pattern recognition, IEEE, pp 335–340
104. Gavrila DM, Davis LS (1996) 3-D model-based tracking of humans in action: a multi-view approach. In: Proceedings cvpr ieee computer society conference on computer vision and pattern recognition, IEEE, pp 73–80
105. Veeraraghavan A, Chellappa R, Roy-Chowdhury AK (2006) The function space of an activity. In: 2006 IEEE Computer society conference on computer vision and pattern recognition (CVPR'06), vol 1: IEEE, pp 959–968
106. Yacoob Y, Black MJ (1999) Parameterized modeling and recognition of activities. Comput Vis Image Underst 73(2):232–247
107. Efros AA, Berg AC, Mori G, Malik J (2003) Recognizing action at a distance. In: Null, IEEE, p 726
108. Lublinerman R, Ozay N, Zarpalas D, Camps O (2006) Activity recognition from silhouettes using linear systems and model (in) validation techniques. In: 18th international conference on pattern recognition (ICPR'06), vol 1: IEEE, pp 347–350
109. Jiang H, Drew MS, Li Z-N (2006) Successive convex matching for action detection. In: 2006 IEEE Computer society conference on computer vision and pattern recognition (CVPR'06), vol 2: IEEE, pp 1646–1653
110. Lin Z, Jiang Z, Davis LS (2009) Recognizing actions by shape-motion prototype trees. In: 2009 IEEE 12th international conference on computer vision, IEEE, pp 444–451
111. Yamato J, Ohya J, Ishii K (1992) Recognizing human action in time-sequential images using hidden markov model. CVPR 92:379–385
112. Starner T, Pentland A (1997) Real-time american sign language recognition from video using hidden Markov models. In: Motion-based recognition, Springer, pp 227–243
113. Vogler C, Metaxas D (1999) Parallel hidden Markov models for American sign language recognition. In: Proceedings of the seventh IEEE international conference on computer vision, vol 1: IEEE, pp 116–122
114. Bobick AF, Wilson AD (1997) A state-based approach to the representation and recognition of gesture. IEEE Trans Pattern Anal Mach Intell 19(12):1325–1337
115. Oliver NM, Rosario B, Pentland AP (2000) A Bayesian computer vision system for modeling human interactions. IEEE Trans Pattern Anal Mach Intell 22(8):831–843
116. Park S, Aggarwal JK (2004) A hierarchical Bayesian network for event recognition of human actions and interactions. Multimedia Syst 10(2):164–179
117. Natarajan P, Nevatia R (2007) Coupled hidden semi markov models for activity recognition. In: 2007 IEEE workshop on motion and video computing (WMVC'07), IEEE, pp 10–10
118. Gupta A, Davis LS (2007) Objects in action: An approach for combining action understanding and object perception. In: 2007 IEEE conference on computer vision and pattern recognition, IEEE, pp 1–8
119. Moore DJ, Essa IA, Hayes MH (1999) Exploiting human actions and object context for recognition tasks. In: Proceedings of the seventh IEEE international conference on computer vision, vol 1: IEEE, pp 80–86
120. Yu E, Aggarwal JK (2009) Human action recognition with extremities as semantic posture representation. In: 2009 IEEE computer society conference on computer vision and pattern recognition workshops, IEEE, pp 1–8
121. Kellokumpu V, Zhao G, Pietikäinen M (2011) Recognition of human actions using texture descriptors. Mach Vis Appl 22(5):767–780
122. Shi Q, Cheng L, Wang L, Smola A (2011) Human action segmentation and recognition using discriminative semi-Markov models. Int J Comput Vision 93(1):22–32
123. Wang L, Suter D (2007) Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model. In: 2007 IEEE conference on computer vision and pattern recognition, IEEE, pp 1–8
124. Rahman SA, Cho S-Y, Leung M (2012) Recognising human actions by analysing negative spaces. IET Comput Vision 6(3):197–213
125. Vishwakarma DK, Kapoor R (2015) Hybrid classifier based human activity recognition using the silhouette and cells. Expert Syst Appl 42(20):6957–6965
126. Junejo IN, Junejo KN, Al Aghbari Z (2014) Silhouette-based human action recognition using SAX-Shapes. The Visual Comput 30(3):259–269
127. Chaaraoui AA, Climent-Pérez P, Flórez-Revuelta F (2013) Silhouette-based human action recognition using sequences of key poses. Pattern Recogn Lett 34(15):1799–1807
128. Chaaraoui AA, Flórez-Revuelta F (2014) A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views. Int Schol Res Notices 2014:6666
129. Cheema S, Eweiwi A, Thurau C, Bauckhage C (2011) Action recognition by learning discriminative key poses. In: 2011 IEEE international conference on computer vision workshops (ICCV Workshops), IEEE, pp 1302–1309
130. Chun S, Lee C-S (2016) Human action recognition using histogram of motion intensity and direction from multiple views. IET Comput Vision 10(4):250–257
131. Murtaza F, Yousaf MH, Velastin SA (2016) Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description. IET Comput Vision 10(7):758–767
132. Ladjailia A, Bouchrika I, Merouani HF, Harrati N, Mahfouf Z (2020) Human activity recognition via optical flow: decomposing activities into basic actions. Neural Comput Appl 32(21):16387–16400
133. Ahmad M, Lee S-W (2006) HMM-based human action recognition using multiview image sequences. In: 18th international conference on pattern recognition (ICPR'06), vol 1: IEEE, pp 263–266
134. Pehlivan S, Forsyth DA (2014) Recognizing activities in multiple views with fusion of frame judgments. Image Vis Comput 32(4):237–249
135. Jiang Z, Lin Z, Davis L (2012) Recognizing human actions by learning and matching shape-motion prototype trees. IEEE Trans Pattern Anal Mach Intell 34(3):533–547
136. Eweiwi A, Cheema S, Thurau C, Bauckhage C (2011) Temporal key poses for human action recognition. In: 2011 IEEE international conference on computer vision workshops (ICCV Workshops), IEEE, pp 1310–1317
137. Shi Y, Huang Y, Minnen D, Bobick A, Essa I (2004) Propagation networks for recognition of partially ordered sequential action. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, CVPR 2004, vol. 2: IEEE, pp II–II
138. Yin J, Meng Y (2010) Human activity recognition in video using a hierarchical probabilistic latent model. In: 2010 IEEE

computer society conference on computer vision and pattern recognition-workshops, IEEE, pp 15–20
139. Mauthner T, Roth PM, Bischof H (2010) Temporal feature weighting for prototype-based action recognition. Asian conference on computer vision. Springer, Berlin, pp 566–579
140. Han L, Wu X, Liang W, Hou G, Jia Y (2010) Discriminative human action recognition in the learned hierarchical manifold space. Image Vis Comput 28(5):836–849
141. Zeng Z, Ji Q (2010) Knowledge based activity recognition with dynamic bayesian network. European conference on computer vision. Springer, Berlin, pp 532–546
142. Minnen D, Essa I, Starner T (2003) Expectation grammars: leveraging high-level expectations for activity recognition. In: 2003 IEEE computer society conference on computer vision and pattern recognition, 2003. Proceedings, vol 2: IEEE, pp II–II
143. Moore D, Essa I (2002) Recognizing multitasked activities from video using stochastic context-free grammar. In: AAAI/IAAI, pp 770–776
144. Kitani KM, Sato Y, Sugimoto A (2008) Recovering the basic structure of human activities from noisy video-based symbol strings. Int J Pattern Recognit Artif Intell 22(08):1621–1646
145. Wang L, Wang Y, Gao W (2011) Mining layered grammar rules for action recognition. Int J Comput Vision 93(2):162–182
146. Nevatia R, Hobbs J, Bolles B (2004) An ontology for video event representation. In: 2004 Conference on computer vision and pattern recognition workshop, IEEE, pp 119–119
147. Ryoo MS, Aggarwal JK (2006) Recognition of composite human activities through context-free grammar based representation. In: 2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06), vol 2: IEEE, pp 1709–1718
148. Pinhanez CS, Bobick AF (1998) Human action detection using pnf propagation of temporal constraints. In: Proceedings. 1998 IEEE computer society conference on computer vision and pattern recognition (Cat. No. 98CB36231), IEEE, pp 898–904
149. Ghanem N, De Menthon D, Doermann D, Davis L (2004) Representation and recognition of events in surveillance video using petri nets. In: 2004 conference on computer vision and pattern recognition workshop, IEEE, pp 112–112
150. Intille SS, Bobick AF (1999) A framework for recognizing multi-agent action from visual evidence. AAAI/IAAI 99(518–525):2
151. Siskind JM (2001) Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. J Artif Intell Res 15:31–90
152. Tran SD, Davis LS (2008) Event modeling and recognition using markov logic networks. European conference on computer vision. Springer, Berlin, pp 610–623
153. Morariu VI, Davis LS (2011) Multi-agent event recognition in structured scenarios. In: CVPR 2011, IEEE, pp 3289–3296
154. Wang H, Schmid C (2013) Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision, pp 3551–3558
155. Kang L, Ye P, Li Y, Doermann D (2014) Convolutional neural networks for no-reference image quality assessment. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1733–1740
156. Banzhaf W, Nordin P, Keller RE, Francone FD (1998) Genetic programming. Springer, Berlin
157. Shao L, Ji L, Liu Y, Zhang J (2012) Human action segmentation and recognition via motion and shape analysis. Pattern Recogn Lett 33(4):438–445
158. Marĉelja S (1980) Mathematical description of the responses of simple cortical cells. JOSA 70(11):1297–1300
159. Primer A, Burrus CS, Gopinath RA (1998) Introduction to wavelets and wavelet transforms. Prentice Hall, Upper Saddle River
160. Harris ZS (1954) Distributional structure. Word 10(2–3):146–162
161. Guha T, Ward RK (2011) Learning sparse representations for human action recognition. IEEE Trans Pattern Anal Mach Intell 34(8):1576–1588
162. Zheng J, Jiang Z, Phillips PJ, Chellappa R (2012) Cross-view action recognition via a transferable dictionary pair. BMVC 1:7
163. Zhu F, Shao L (2014) Weakly-supervised cross-domain dictionary learning for visual recognition. Int J Comput Vision 109(1–2):42–59
164. Kim H-J, Lee JS, Yang H-S (2007) Human action recognition using a modified convolutional neural network. International symposium on neural networks. Springer, Berlin, pp 715–723
165. Jones JP, Palmer LA (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J Neurophysiol 58(6):1233–1258
166. Kim H-J, Lee J, Yang H-S (2006) A weighted FMM neural network and its application to face detection. International conference on neural information processing. Springer, Berlin, pp 177–186
167. Jhuang H, Serre T, Wolf L, Poggio T (2007) A biologically inspired system for action recognition. In: 2007 IEEE 11th international conference on computer vision, IEEE, pp 1–8
168. Shao L, Liu L, Li X (2013) Feature learning for image classification via multiobjective genetic programming. IEEE Trans Neural Netw Learn Syst 25(7):1359–1371
169. Taylor GW, Hinton GE, Roweis ST (2007) Modeling human motion using binary latent variables. In: Advances in neural information processing systems, pp 1345–1352
170. Baum LE, Petrie T (1966) Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Stat 37(6):1554–1563
171. Ji S, Xu W, Yang M, Yu K (2012) 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
172. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
173. Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR 2011, IEEE, pp 3361–3368
174. Hyvarinen A, Hurri J, Hoyer PO (2009) "A probabilistic approach to early computational vision. Nat Image Stat 2:666
175. Wold S, Esbensen K, Geladi P (1987) Principal component analysis. Chemom Intell Lab Syst 2(1–3):37–52
176. Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2011) Sequential deep learning for human action recognition. International workshop on human behavior understanding. Springer, Berlin, pp 29–39
177. Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229
178. Jia Y et al. (2014) Caffe: convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, ACM, pp 675–678
179. Ning F, Delhomme D, LeCun Y, Piano F, Bottou L, Barbano PE (2005) Toward automatic phenotyping of developing embryos from videos. IEEE Trans Image Process 14(9):1360–1371
180. Singh T, Vishwakarma DK (2021) A deeply coupled ConvNet for human activity recognition using dynamic and RGB images. Neural Comput Appl 33(1):469–485

181. Yao L, Qian Y (2018) Dt-3dresnet-lstm: An architecture for temporal activity recognition in videos. Pacific Rim conference on multimedia. Springer, Berlin, pp 622–632
182. Meng B, Liu X, Wang X (2018) Human action recognition based on quaternion spatial-temporal convolutional neural network and LSTM in RGB videos. Multimedia Tools Appl 77(20):26901–26918
183. Qi M, Qin J, Li A, Wang Y, Luo J, Van Gool L (2018) stagnet: an attentive semantic RNN for group activity recognition. In: Proceedings of the European conference on computer vision (ECCV), pp 101–117
184. Qi M, Wang Y, Qin J, Li A, Luo J, Van Gool L (2019) stagNet: an attentive semantic RNN for group activity and individual action recognition. IEEE Trans Circuits Syst Video Technol 30(2):549–565
185. Muhammad K et al (2021) Human action recognition using attention based LSTM network with dilated CNN features. Futur Gener Comput Syst 125:820–830
186. He J-Y, Wu X, Cheng Z-Q, Yuan Z, Jiang Y-G (2021) DB-LSTM: Densely-connected Bi-directional LSTM for human action recognition. Neurocomputing 444:319–331
187. Hu K, Zheng F, Weng L, Ding Y, Jin J (2021) Action recognition algorithm of Spatio-temporal differential LSTM based on feature enhancement. Appl Sci 11(17):7876
188. Vaswani A et al. (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
189. Neimark D, Bar O, Zohar M, Asselmann D (2021) Video transformer network. arXiv preprint arXiv:2102.00719
190. Plizzari C, Cannici M, Matteucci M (2021) Spatial temporal transformer network for skeleton-based action recognition. International conference on pattern recognition. Springer, Berlin, pp 694–701
191. Mazzia V, Angarano S, Salvetti F, Angelini F, Chiaberge M (2021) Action transformer: a self-attention model for short-time human action recognition. arXiv preprint arXiv:2107.00606
192. Ullah A, Muhammad K, Haq IU, Baik SW (2019) Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments. Futur Gener Comput Syst 96:386–397
193. Chong YS, Tay YH (2017) Abnormal event detection in videos using spatiotemporal autoencoder. International symposium on neural networks. Springer, Berlin, pp 189–196
194. Cui R, Hua G, Wu J (2020) AP-GAN: predicting skeletal activity to improve early activity recognition. J Vis Commun Image Represent 73:102923
195. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
196. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4305–4314
197. Sánchez J, Perronnin F, Mensink T, Verbeek J (2013) Image classification with the fisher vector: theory and practice. Int J Comput Vision 105(3):222–245
198. Gowda SN, Sevilla-Lara L, Keller F, Rohrbach M (2021) CLASTER: clustering with reinforcement learning for zero-shot action recognition. arXiv preprint arXiv:2101.07042
199. Liu K, Liu W, Ma H, Huang W, Dong X (2019) Generalized zero-shot learning for action recognition with web-scale video data. World Wide Web 22(2):807–824
200. Ornek EP (2020) Zero-shot activity recognition with videos. arXiv preprint arXiv:2002.02265
201. Taylor GW, Fergus R, LeCun Y, Bregler C (2010) Convolutional learning of spatio-temporal features. European conference on computer vision. Springer, Berlin, pp 140–153
202. Collobert R, Weston J (2008) A unified architecture for natural language processing: Deep neural networks with multitask learning. In: Proceedings of the 25th international conference on Machine learning, pp 160–167
203. Yan Y, Ricci E, Subramanian R, Liu G, Sebe N (2014) Multitask linear discriminant analysis for view invariant action recognition. IEEE Trans Image Process 23(12):5599–5611
204. Yang Q (2009) Activity recognition: linking low-level sensors to high-level intelligence. In: Twenty-first international joint conference on artificial intelligence
205. Zheng VW, Hu DH, Yang Q (2009) Cross-domain activity recognition. In: Proceedings of the 11th international conference on Ubiquitous computing, pp 61–70
206. Liu J, Shah M, Kuipers B, Savarese S (2011) Cross-view action recognition via view knowledge transfer. In: CVPR 2011, IEEE, pp 3209–3216
207. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1717–1724
208. Wang H, Schmid AC, Liu C-L (2011) Action recognition by dense trajectories. Proc IEEE Conf Comput Vis Pattern Recognit 2:3169–3176
209. Kliper-Gross O, Gurovich Y, Hassner T, Wolf L (2012) Motion interchange patterns for action recognition in unconstrained videos. European conference on computer vision. Springer, Berlin, pp 256–269
210. Oneata D, Verbeek J, Schmid C (2013) Action and event recognition with fisher vectors on a compact feature set. In: Proceedings of the IEEE international conference on computer vision, pp 1817–1824
211. Jain M, Jégou H, Bouthemy P (2013) Better exploiting motion for better action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2555–2562
212. Peng X, Zou C, Qiao Y, Peng Q (2014) Action recognition with stacked fisher vectors. European conference on computer vision. Springer, Berlin, pp 581–595
213. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. arXiv preprint arXiv:1406.2199
214. Sun L, Jia K, Yeung D-Y, Shi BE (2015) Human action recognition using factorized spatio-temporal convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4597–4605
215. Wang L, Xiong Y, Wang Z, Qiao Y (2015) Towards good practices for very deep two-stream convnets. arXiv preprint arXiv:1507.02159
216. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4694–4702
217. Fernando B, Gavves E, Oramas JM, Ghodrati A, Tuytelaars T (2015) Modeling video evolution for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5378–5387
218. Donahue J et al. (2015) Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2625–2634
219. Jiang Y-G, Dai Q, Liu W, Xue X, Ngo C-W (2015) Human action recognition in unconstrained videos by explicit motion modeling. IEEE Trans Image Process 24(11):3781–3795
220. Lan Z, Lin M, Li X, Hauptmann AG, Raj B (2015) Beyond gaussian pyramid: Multi-skip feature stacking for action

recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 204–212

221. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
222. Fernando B, Gould S (2016) Learning end-to-end video classification with rank-pooling. In: International conference on machine learning, PMLR, pp 1187–1196
223. Fernando B, Anderson P, Hutter M, Gould S (2016) Discriminative hierarchical rank pooling for activity recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1924–1932
224. Li Y, Li W, Mahadevan V, Vasconcelos N (2016) Vlad3: encoding dynamics of deep features for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1951–1960
225. Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1933–1941
226. Varol G, Laptev I, Schmid C (2017) Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1510–1517
227. Singh D, Mohan CK (2017) Graph formulation of video activities for abnormal activity recognition. Pattern Recogn 65:265–272
228. Carmona JM, Climent J (2018) Human action recognition by means of subtensor projections and dense trajectories. Pattern Recogn 81:443–455
229. Mao F, Wu X, Xue H, Zhang R (2018) Hierarchical video frame sequence representation with deep convolutional graph network. In: Proceedings of the European conference on computer vision (ECCV) workshops, pp 0–0
230. Siddiqi MH, Alruwaili M, Ali A (2019) A novel feature selection method for video-based human activity recognition systems. IEEE Access 7:119593–119602
231. Zhang Y, Po LM, Liu M, Rehman YAU, Ou W, Zhao Y (2020) Data-level information enhancement: motion-patch-based Siamese convolutional neural networks for human activity recognition in videos. Expert Syst Appl 147:113203
232. Arzani MM, Fathy M, Azirani AA, Adeli E (2020) Switching structured prediction for simple and complex human activity recognition. IEEE Trans Cybern 6:7777
233. Gowda SN, Rohrbach M, Sevilla-Lara L (2020) SMART frame selection for action recognition. arXiv e-prints, p. arXiv:2012.10671
234. Wharton Z, Behera A, Liu Y, Bessis N (2021) Coarse temporal attention network (cta-net) for driver's activity recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1279–1289
235. Ullah A, Muhammad K, Ding W, Palade V, Haq IU, Baik SW (2021) Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications. Appl Soft Comput 103:107102
236. Khan MA et al (2021) A fused heterogeneous deep neural network and robust feature selection framework for human actions recognition. Arabian J Sci Eng 6:1–16
237. Ullah A, Muhammad K, Hussain T, Baik SW (2021) Conflux LSTMs network: a novel approach for multi-view action recognition. Neurocomputing 435:321–329
238. Reinolds F, Neto C, Machado J (2022) Deep learning for activity recognition using audio and video. Electronics 11(5):782
239. Siddiqi MH, Alsirhani A (2022) An efficient feature selection method for video-based activity recognition systems. Math Problems Eng 2022:66689
240. Khare M, Jeon M (2022) Multi-resolution approach to human activity recognition in video sequence based on combination of complex wavelet transform, Local Binary Pattern and Zernike moment. Multimedia Tools Appl 2:1–30
241. Deotale D et al (2022) HARTIV: human activity recognition using temporal information in videos. CMC-Comput Mater Continua 70(2):3919–3938
242. Zhang C, Wu J, Li Y (2022) ActionFormer: localizing moments of actions with transformers. arXiv preprint arXiv:2202.07925
243. Ahmed N, Asif HMS, Khalid H (2021) PIQI: perceptual image quality index based on ensemble of Gaussian process regression. Multimedia Tools Appl 80(10):15677–15700
244. Ahmed SAN (2022) BIQ2021: a large-scale blind image quality assessment database. arXiv preprint arXiv:submit/4155160
245. Ahmed N, Asif HS, Bhatti AR, Khan A (2022) Deep ensembling for perceptual image quality assessment. Soft Comput 2:1–22
246. Ahmed N, Asif HMS (2020) Perceptual quality assessment of digital images using deep features. Comput Inform 39(3):385–409
247. Alzantot M, Chakraborty S, Srivastava M (2017) Sensegen: a deep learning architecture for synthetic sensor data generation. In: 2017 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops), IEEE, pp 188–193